

基于特征优化和LS-SVM的棉田杂草识别*

李先锋^{1,2} 朱伟兴¹ 纪滨¹ 刘波¹

(1. 江苏大学电气信息工程学院, 镇江 212013; 2. 盐城工学院信息工程学院, 盐城 224051)

【摘要】 为了提高杂草识别的精度和效率,提出了一种基于特征优化和最小二乘支持向量机(LS-SVM)技术的棉田杂草识别方法。在对原始图像灰度化、滤波去噪和阈值分割等处理的基础上,提取植物叶片的6个几何特征和7个Hu不变矩,用粒子群优化(PSO)算法对形状特征进行优化选择,缩减LS-SVM训练样本数据,然后用训练好的分类器进行杂草识别。实验结果表明,该方法在有效缩减形状特征的同时,能够保持高于原始特征集的识别率,平均正确识别率达到95.8%。

关键词: 棉花 杂草识别 特征选择 粒子群优化算法 最小二乘支持向量机

中图分类号: TP391.41 **文献标识码:** A **文章编号:** 1000-1298(2010)11-0168-05

Weed Identification Based on Features Optimization and LS-SVM in the Cotton Field

Li Xianfeng^{1,2} Zhu Weixing¹ Ji Bin¹ Liu Bo¹

(1. School of Electronic & Information Engineering, Jiangsu University, Zhenjiang 212013, China

2. School of Information Engineering, Yancheng Institute of Technology, Yancheng 224051, China)

Abstract

In order to improve the accuracy and efficiency of weed identification, a method for cotton-weed recognition was proposed by using the combination technique of features optimization and least squares support vector machine (LS-SVM). After a series of image processing such as graying, filtering and threshold segmenting, six geometric shape features and seven Hu moment invariants were extracted from the single plant leaf. Then, using particle swarm optimization (PSO) algorithm, the extracted features were optimized in order to reduce the size of the training data sets. Finally, the weed was identified by using the trained classifier. The experimental results indicate that this method can effectively compact feature subset and maintain a higher accuracy than using the original feature set, the average correct identification rate is 95.8%.

Key words Cotton, Weed identification, Feature selection, Particle swarm optimization, Least squares support vector machine

引言

利用计算机视觉和图像处理技术将杂草从背景中识别进行准确定位和变量喷洒农药已成为现代除草技术研究的主要方向^[1]。杂草识别是指通过提

取和选择杂草及作物图像的视觉特征(形状、颜色、纹理等)将杂草从作物中分离出来,其本质是一种分类问题。要实现精确识别,就要尽可能引入多种特征,而特征空间的复杂程度必然会提高。因此,在众多特征中如何选取最有效的特征变量,尽量减少

收稿日期: 2010-01-26 修回日期: 2010-03-30

* 江苏大学现代农业装备与技术省部共建教育部重点实验室开放基金资助项目(NZ200709)

作者简介: 李先锋,博士生,盐城工学院讲师,主要从事图像处理与模式识别研究,E-mail: vanguardlee@sohu.com

通讯作者: 朱伟兴,教授,博士生导师,主要从事智能检测、图像处理与机器视觉研究,E-mail: wxzhu@ujs.edu.cn

计算量,是能否实现快速、实时、准确识别杂草的关键。这就要求在对杂草和作物进行图像处理的基础上,利用智能算法、组合优化等理论进行特征的优化选择,找到具有较好可分性的特征子集,从而降低分类器输入维数和学习复杂度,提高分类识别的精度和效率^[2~3]。

本文运用数字图像处理技术对实地拍摄到的棉花和杂草原始图像进行一系列处理,有效获取棉花和杂草图像的 shape 特征参数,然后用粒子群优化 (particle swarm optimization, 简称 PSO) 算法对特征进行优化,选出最佳适应度的特征子集对最小二乘支持向量机 (least squares support vector machine, 简称 LS-SVM) 分类器进行训练,用训练后的 LS-SVM 进行杂草和作物的分类识别。

1 图像采集及预处理

在江苏盐城大冈水旱轮作棉田区,选取不同时段在自然光源下由数码相机拍摄的棉花和杂草图像 (马唐、稗、小酸浆、千金子、马齿苋、刺儿菜、铁苋菜等),并存储在棉田杂草图库中用于实验。

如图 1a 所示,以杂草小酸浆的原始图像为例 (其他杂草及棉花类似),为了获取更多的信息量,首先需要将原始彩色图像灰度化,然后进行滤波去噪处理,最后进行阈值分割,得到二值化图像。

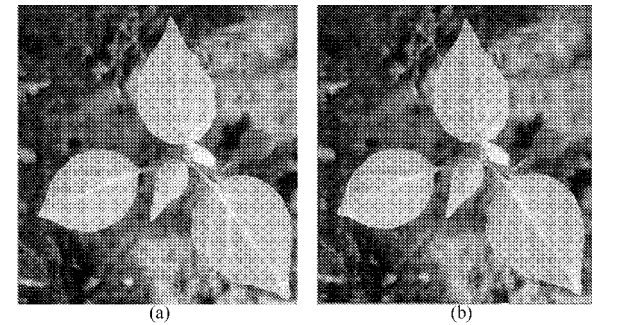


图 1 小酸浆图像灰度化

Fig. 1 Gray image of physalis minima coloring picture
(a) 原始图像 (b) 超绿色法灰度图像

图像灰度化处理:由于植物的反射光谱特性不同于土壤背景,故自然光线下获取的杂草和棉花图像中背景与植物呈现明显的区别。本文采用 Woebbecke 等提出的超绿色 (excess green) 方法^[4]进行彩色图像灰度化处理。由三基色原理可知,彩色图像的 R 、 G 、 B 三色分量中植物的 G 值与背景的 G 值存在很大差异,因此可以由 R 、 G 、 B 三个子图得到一幅有效图像,公式为

$$E_G(x,y) = 2G(x,y) - R(x,y) - B(x,y) \quad (1)$$

式中 $G(x,y)$ 、 $R(x,y)$ 、 $B(x,y)$ ——绿、红、蓝三色分量矩阵

$E_G(x,y)$ ——超绿色值

根据式 (1) 对图 1a 所示的杂草小酸浆图像进行灰度化处理,结果如图 1b 所示。

灰度图像滤波处理:为了消除混杂在图像中的干扰,改善图像质量,强化图像表现特征,需要对灰度图像进行滤波处理。本文选择算法相对简单、有利于实时性的邻域均值滤波法和中值滤波法。图 2 是运用 3×3 窗口对图 1b 灰度图像分别进行邻域滤波和中值滤波处理得到的结果。从滤波效果看,两者差别不大,都能有效滤除噪声;但从处理速度及编程复杂程度看,邻域平均滤波有明显优势,兼顾了滤波质量和实时性要求。

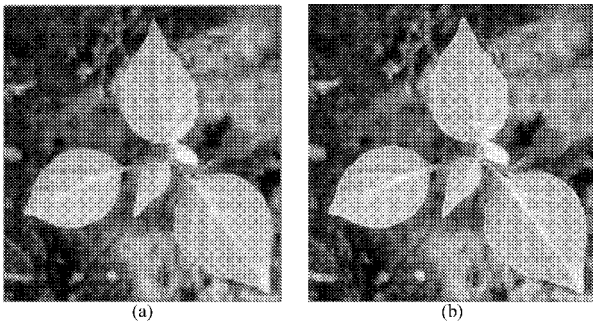


图 2 灰度图的滤波处理

Fig. 2 Filter results of the weed's gray image
(a) 中值滤波 (b) 邻域滤波

图像二值化处理:为了便于其后进行形状特征的提取和方便计算叶片的形状特征参数,本文通过图像工具截取经滤波的杂草灰度图像中单个叶片的区域,然后对该区域图像实施阈值化分割,采用迭代阈值选择法来计算灰度阈值,把叶片与其背景分开并形成二值图像。具体步骤为:①求出整幅图像的最大灰度值 $T_{\max}$ 和最小灰度值 $T_{\min}$,令初始灰度阈值 $T_0 = (T_{\max} + T_{\min})/2$ 。②根据灰度阈值 T_k 将图像分割为前景和背景,分别求出二者的平均灰度值 M_1 和 M_2 。③求出新灰度阈值 $T_{k+1} = (M_1 + M_2)/2$ 。④若 $T_{k+1} = T_k$,则 T_k 即为灰度阈值,否则令 $k = k + 1$,转至步骤②继续迭代计算。如图 3a 为截取图 1a 中小酸浆的一个单叶图像,经上述一系列图像处理得

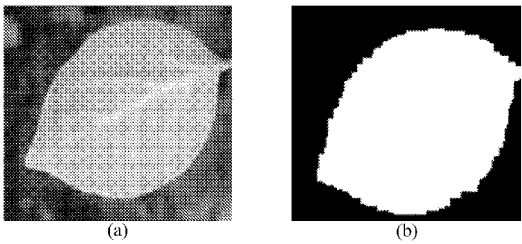


图 3 单叶阈值分割结果

Fig. 3 Threshold segmentation result of weed's single leaf image
(a) 杂草单叶图像 (b) 单叶二值图像

到单叶二值图像,如图 3b 所示。

2 形状特征提取

形状特征是区分杂草和作物的主要特征之一,并在已有杂草识别研究中取得了一定的效果,特别是神经网络、支持向量机、分形等数学工具的引入进一步提高了对杂草的识别率^[5~7]。本文以植物叶片形状为特征,对图 3b 所示的小酸浆单叶二值图像采

用基于递归算法的区域标记方法^[8]获得叶片的面积、周长、长度、宽度等几个简单的区域形状特征,并以此计算出复杂度、宽长比、离散度、圆度、伸长度、矩形度等 6 个具有 RST(旋转、比例、平移)不变性的无量纲几何特征参数。

表 1 为棉田植物叶片几何形状特征定义及杂草小酸浆 4 个叶片中圆度为中间值的叶片几何特征参数值。

表 1 植物叶片形状特征定义及小酸浆叶片特征参数值

Tab.1 Definition of leaf shape character and parameters of physalis minima leaf

特征参数	参数含义	符号或计算式	数值
面积	叶片图像区域中的像素点数	A	808
周长	叶片图像区域边缘像素点数	P	115
宽度	叶片最小外接矩宽(像素点数)	W	28
长度	叶片最小外接矩长(像素点数)	L	35
复杂度	叶片区域的形状复杂程度	$C = 4\pi A/P^2$	0.768
宽长比	叶片区域最小外接矩的宽长比	$K = W/L$	0.8
离散度	叶片区域单位面积的周长	$e = P^2/A$	16.37
圆度	叶片区域与外接圆紧凑度关联程度	$D = 4\pi A/L^2$	8.28
伸长度	叶片区域细长程度	$E = (L - W)/(L + W)$	0.111
矩形度	叶片区域对其外接矩的充满程度	$R = A/(LW)$	0.82

考虑到 4 个区域形状特征会随叶片生长而发生明显变化,可比性小,故不宜作为杂草识别的特征依据,而应该利用与叶片大小、方向等无关的规范化后的单个或多个无量纲参数的组合来提高识别效果。

另外,除了几何特征外,图像矩也因其具有的 RST 不变性而被广泛应用于图像识别领域。本文加入 Hu M K 提出的 7 项 Hu 不变矩^[9]作为识别特征,并采用 Chen Chaur-Chin 提出的改进矩算法^[10],利用 OpenCV 中的函数 cvGetHumoments() 计算这 7 项 Hu 不变矩。图 3b 中叶片的 7 项 Hu 矩数值分别为: $1.834\,72 \times 10^{-1}$ 、 $5.903\,38 \times 10^{-3}$ 、 $1.425\,81 \times 10^{-4}$ 、 $3.827\,06 \times 10^{-5}$ 、 $2.834\,91 \times 10^{-9}$ 、 $2.920\,53 \times 10^{-6}$ 、 $1.365\,70 \times 10^{-11}$ 。对于每个叶片,采用其 6 项几何特征和 7 项 Hu 不变矩共 13 项 RST 不变性特征作为分类特征。

3 特征选择和 LS - SVM 分类器参数同步优化算法

相比传统的统计方法、神经网络和其他分类算法,SVM 通过结构风险最小化原理提高了泛化能力,并且具有很好的学习性能,较好地解决了小样本、非线性、高维数等问题,已逐渐成为解决模式分类问题的首选工具^[11]。LS - SVM 是基于 SVM 的一种改进算法,它把 SVM 中解决二次规划问题的不等式约束转换成了等式约束,使得原 SVM 算法的二次

寻优变为难度更低、速度更快的线性方程求解。因此,本文选择 LS - SVM^[12]作为分类器模型,描述如下:

给定训练集 $\{x_k, y_k\}_{k=1}^N$, LS - SVM 定义优化问题 $\min_{w,b,\xi} J(w, \xi) = \frac{1}{2} w^T w + \gamma \frac{1}{2} \sum_{i=1}^N \xi_i^2$, 满足约束 $y_i = w^T \phi(x_i) + b + \xi_i, i = 1, 2, \dots, l$, 其中 $x \in \mathbf{R}^n, y \in \mathbf{R}$, 非线性函数 $\phi(\cdot): \mathbf{R}^n \rightarrow \mathbf{R}$ 将输入空间映射为高维特征空间。相应的 Lagrange 函数为 $L = J - \sum_{i=1}^l \alpha_i (w^T \phi(x_i) + b + \xi_i - y_i), i = 1, 2, \dots, l$, 该优化问题转化为求

$$\begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 & I^T \\ I & K + \gamma^T I \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ y \end{bmatrix}$$

(2)

其中 $\alpha = [\alpha_1, \alpha_2, \dots, \alpha_N]^T, I = [1, 1, \dots, 1]^T, y = [y_1, y_2, \dots, y_N]^T; K$ 为方阵,其第 m 列第 l 行的元素是 $K(x_m, x_l)$,即核函数。

求解上述问题后得到的最优分类函数,即 LS - SVM 构造出的最优分类超平面为

$$y(x) = \text{sgn} \left(\sum_{k=1}^N a_k y_k K(x, x_k) + b \right)$$

(3)

本文核函数采用径向基函数 RBF: $K(x_i, x_j) = \exp(-\|x_i - x_j\|^2 / (2\sigma^2))$ 。

为了降低分类器学习和训练的时间及空间复杂度,提高分类的精度和速度,需要用特征选择算法找

到具有较好可分性的特征子空间;同时,上述分类器惩罚参数 γ 和核参数 σ 在很大程度上决定了 LS-SVM 的学习和泛化能力。为此,本文提出了一种基于 PSO 的特征选择及 LS-SVM 参数同步优化算法。PSO 是一种基于群体智能理论的随机寻优进化算法^[13],其基本思想是:系统初始化为一组随机的可行解,通过迭代在可行解空间搜索最优值。在每次迭代中,首先按一定规则对粒子 i 的速度 v_i 和位置 x_i 进行更新;然后求各粒子的适应度(Fitness),计算该粒子当前的最优值,即局部最优解 p_{best} ,计算整个种群当前的最优值,即全局最优解 g_{best} 。迭代过程中粒子的速度和位置更新公式为

$$v_{i+1} = \omega v_i + c_1 r_1 (p_{\text{best}} - x_i) + c_2 r_2 (g_{\text{best}} - x_i) \tag{4}$$

$$x_{i+1} = x_i + (k + \text{rand}[\]) v_{i+1} \tag{5}$$

如果满足终止条件,则结束迭代。式中 c_1 、 c_2 表示加速常数, r_1 、 r_2 表示随机数。

采用适应度函数来评价粒子的优劣性能,度量 LS-SVM 的泛化能力,即

$$F(i) = A(i)/(1 + \lambda n(i)) \tag{6}$$

式中 $F(i)$ ——粒子生成解的适应度
 $A(i)$ ——粒子在 LS-SVM 中的分类精度
 $n(i)$ ——该粒子选中的变量数(特征个数)
 λ ——特征个数的权重参数,用来平衡最大正分率与变量个数

由式(6)知,粒子使分类器产生的分类精度越高,选出的特征数目就越少, F 就越大。

特征选择及分类器参数同步优化算法的实现步骤如下:

- (1)初始化粒子群 $X(t)$ 、PSO 参数、最大迭代次数及 LS-SVM 参数,输入样本数据。
- (2)测试并记录分类精度,按式(6)计算各粒子的适应度 $F(i)$ 并更新局部最优解 p_{best} 和全局最优解 g_{best} :if $F(x_i) > F(p_{\text{best}})$, then $p_{\text{best}} = x_i$; if $F(x_i) > F(g_{\text{best}})$, then $g_{\text{best}} = x_i$ 。
- (3)根据式(4)和式(5),更新粒子的速度和位置,产生新种群 $X(t+1)$ 。
- (4)判断是否满足中止条件(if $t < T_{\text{max}}$),如满足则结束寻优,输出当前最优的特征子集、LS-SVM 参数 γ 和 σ 、分类精度等,此时的分类器即为训练好的 LS-SVM 分类器;否则令 $t = t + 1$,跳转到步骤(2),继续新一轮寻优。

4 杂草识别实验分析

4.1 实验环境及数据准备

实验在主频 3.0 GHz 奔腾 IV、内存 1 GB、Windows XP 操作系统的 PC 机上进行,实验程序用

Matlab 编制(Matlab 7.6.5 软件平台),分类器选用 LS-SVM lab1.5。

本实验拟采用的形状特征包括: x_1 复杂度、 x_2 宽长比、 x_3 离散度、 x_4 圆度、 x_5 伸长度、 x_6 矩形度、 x_7 Hu-1、 x_8 Hu-2、 x_9 Hu-3、 x_{10} Hu-4、 x_{11} Hu-5、 x_{12} Hu-6 和 x_{13} Hu-7。每次实验随机选择从棉田采集的 70 幅原始图像(杂草或棉花),其中 30 幅用作训练,40 幅用作测试,按上文的方法进行图像处理和特征提取,并分别计算 13 个形状特征参数。

4.2 特征选择及识别结果分析

(1)形状特征归一化处理

从表 1 特征参数值和 Hu 不变矩的数据可以看出,13 个形状特征参数量级相差较大,为避免出现计算饱和现象,对数据进行 $[0,1]$ 归一化处理,即

$$E' = (E - E_{\text{min}})/(E_{\text{max}} - E_{\text{min}}) \tag{7}$$

式中 E 、 E' ——转换前、后的数据
 E_{max} 、 E_{min} ——每类样本各特征参数的最大值、最小值

(2)分类器训练及识别结果分析

分别以 6 个几何特征、所有 13 个形状特征为两组特征集,利用以下两种方法进行 3 次杂草识别实验。方法一采用 LS-SVM 方法,不经过特征优化,分类器最优参数根据经验在一定范围内间隔选取确定。方法二采用基于 PSO 的 LS-SVM 方法,同步进行特征选择和分类器参数自动选优:设粒子群规模 $m = 30$,最大迭代次数 $T_{\text{max}} = 40$,初始惯性权重因子 $\omega_{\text{max}} = 0.9$ (线性下降到 $\omega_{\text{min}} = 0.2$),加速常数 $c_1 = c_2 = 2.05$, r_1 、 $r_2 \in \text{rand}[0,1]$,调节系数 $k = 0.1$,粒子速度初值 $v_i \in \text{rand}[-5,5]$;分类器参数 $\gamma = \text{rand}[0.01,100]$, $\sigma^2 = \text{rand}[0.01,10]$ 。具体操作如下:首先输入特征集,用训练样本对两种分类方法分别进行训练,然后用训练后的模型对测试数据进行分类识别,结果如表 2 所示。

分析表 2 可知:用于分类的最优特征子集为 $\{x_1, x_2, x_3, x_4, x_7, x_8\}$,包含 4 个几何特征(复杂度、宽长比、离散度、圆度)和 2 个 Hu 不变矩(Hu-1、Hu-2);特征选择后,特征数目明显缩减(13 个减少为 6 个),同时能保持高于原始特征集的识别率,平均分类正确率达到 95.8%;优化算法在寻优过程中保留了对分类贡献较大的特征,实现了特征的优化组合;伸长度、矩形度以及不变矩 Hu-3 ~ Hu-7 等 7 个不良特征参与分类不但增加运算规模,还会给分类器带来干扰,影响分类效果;仅用几何特征,识别率就可达到 90%,且最优特征子集中几何特征也占较大比重,说明大部分杂草通过几何特征就可识别。

表 2 棉田杂草特征选择及识别结果

Tab.2 Experimental results of weed’s features selection and recognition

实验序号	LS - SVM		基于 PSO 的 LS - SVM	
	特征选择结果(特征子集/特征数)	正确识别率/%	特征选择结果(特征子集/特征数)	正确识别率/%
1	$\{x_1, x_2, \cdots, x_6\}/6$	90.0	$\{x_1, x_2, x_3\}/3$	92.5
	$\{x_1, x_2, \cdots, x_{13}\}/13$	92.5	$\{x_1, x_2, x_3, x_4, x_7, x_8\}/6$	95.0
2	$\{x_1, x_2, \cdots, x_6\}/6$	92.5	$\{x_1, x_2, x_3, x_4\}/4$	95.0
	$\{x_1, x_2, \cdots, x_{13}\}/13$	92.5	$\{x_1, x_2, x_3, x_4, x_7, x_8\}/6$	97.5
3	$\{x_1, x_2, \cdots, x_6\}/6$	90.0	$\{x_1, x_2, x_3, x_4\}/4$	92.5
	$\{x_1, x_2, \cdots, x_{13}\}/13$	92.5	$\{x_1, x_2, x_3, x_4, x_7, x_8\}/6$	95.0

另外,杂草与作物之间可能存在叶片交叠(或遮挡)现象,使得叶片的形状特征参数计算不准确,也是影响识别准确率的原因之一。

5 结束语

本文提出了一种基于特征优化的 LS - SVM 杂草分类识别方法。通过图像处理,提取了 13 个适合

反映棉花和杂草形态结构的特征参数,利用 PSO 算法快速寻优的特点,同步进行杂草的特征选择和 LS - SVM分类器的参数优化。棉田杂草识别实验结果表明,本方法在实现杂草形状特征参数有效缩减的同时,平均正确识别率达到 95.8%,说明利用形状特征,使用基于 PSO 的 LS - SVM 技术识别棉田杂草是可行的。

参 考 文 献

1 Naeem A M, Irshad Ahmad, Muham Mad Islam, et al. Weed classification using angular cross sectional intensities for real-time selective herbicide applications [C] // Proceedings of the International Conference on Computing: Theory and Applications, 2007: 70 ~ 74.

2 罗瑜, 易文德, 王丹琛. 大规模数据集下支持向量机训练样本的缩减策略[J]. 计算机科学, 2007, 34(10): 211 ~ 213.

3 Liu Huan, Yu Lei. Toward integrating feature selection algorithms for classification and clustering[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(5): 491 ~ 502.

4 Woebbecke D M, Meyer G E. Color indices for weed identification under various soil, residue, and lighting conditions[J]. Transactions of the ASAE, 1995, 38(1): 259 ~ 269.

5 袁海波, 史岩. 基于神经网络的杂草识别试验研究[J]. 机电工程, 2008, 25(4): 28 ~ 30.

6 吴兰兰, 刘俭英, 文友先, 等. 基于支持向量机的玉米田间杂草识别方法[J]. 农业机械学报, 2009, 40(1): 162 ~ 166. Wu Lanlan, Liu Jianying, Wen Youxian, et al. Weed identification method based on SVM in the corn field[J]. Transactions of the Chinese Society for Agricultural Machinery, 2009, 40(1): 162 ~ 166. (in Chinese)

7 吴兰兰, 刘俭英, 文友先. 基于分形维数的玉米和杂草图像识别[J]. 农业机械学报, 2009, 40(3): 176 ~ 179. Wu Lanlan, Liu Jianying, Wen Youxian. Image identification of corn and weed based on fractal dimension[J]. Transactions of the Chinese Society for Agricultural Machinery, 2009, 40(3): 176 ~ 179. (in Chinese)

8 宋培华, 高教岳. 边界跟踪与序贯算法在二值连通区域标记中的应用[J]. 计算机科学, 2002, 29(3): 108 ~ 110.

9 Hu M K. Visual pattern recognition by moment invariants[J]. IRE Transaction Information Theory, 1962, 8(2): 179 ~ 187.

10 Chen Chaur-Chin. Improved movement invariants for shape discrimination[J]. Pattern Recognition, 1993, 26(5): 683 ~ 686.

11 Zhang Xuegong. Introduction to statistical learning theory and support vector machines[J]. Acta Automatica Sinica, 2000, 26(1): 32 ~ 42.

12 Suykens J A K, Branbanter J K, Lukas L, et al. Weighted least squares support vector machines: robustness and spare approximation[J]. Neurocomputing, 2002, 48(1): 85 ~ 105.

13 Kennedy J. The particle swarm: social adaptation of knowledge[C] // Proceedings of IEEE International Conference on Evolutionary Computation, Indianapolis, Indiana, USA, 1997.