

基于加权二叉树的自适应遗传算法研究

李龙龙¹, 王美丽²

(1. 陕西工业职业技术学院, 陕西 咸阳 712000;

2. 英国伯恩茅斯大学媒体学院, 英国 普尔市 BH92JF)

摘要:为改进传统遗传算法局部搜索能力较差、收敛速度慢等缺点,提出一种基于加权二叉树的遗传算法。通过构建遗传基因二叉树,对种群染色体进行编码,根据子代基因的适应值挑选优秀基因替换弱势基因,采用蚁群信息素对不同的遗传基因进行加权操作,依权重择优进行交叉操作,利用自适应排序选择最优解,并通过对比实验对该算法和基本遗传算法进行了全方位的比较。试验结果表明该算法大大提高了遗传算法的局部搜索能力,加快了算法的收敛速度。

关键词:加权二叉树;自适应遗传算法;蚁群算法;排序选择

中图分类号:TP311

文献标识码:A

文章编号:1673-629X(2010)11-0095-05

Research on an Adaptive Genetic Algorithm Based on Weighted Binary Tree

LI Long-long¹, WANG Mei-li²

(1. Shaanxi Polytechnic Institute, Xianyang 712000, China;

2. Bournemouth University, Poole BH92JF, UK)

Abstract: In order to improve the deficiencies of the locally searching capability and slow convergence speed of traditional genetic algorithm, a novel genetic algorithm based on weighted binary tree is presented in this paper. Constructing the genetic binary tree by adopting the pheromone of ant colony to weight different genes and a new crossover strategy which chooses the outstanding individuals according to weight is introduced, finally an adaptive adjusting mechanism is adopted by the ranking selection strategy to choose the best solution. By comparing the algorithm and the basic genetic algorithm, the experimental results showed that the algorithm greatly improved the local search capability and convergence speed of the genetic algorithm.

Key words: weighted binary tree; adaptive genetic algorithm; ant colony algorithm; ranking selection

0 引言

20世纪90年代初,意大利人 M. Dorigo 等率先提出了一种元启发式的搜索算法,即蚁群算法^[1,2]。该算法是通过模拟蚂蚁的信息激素交互机制来选择新的路径,进而达到寻找最优解的目的。而遗传算法是美国的 J. Holland 教授^[3,4]提出的一种根据生物界“物竞天择、适者生存”的进化机制发展起来的具有鲁棒性、自适应性以及自我组织能力的全局优化随机搜索算法,在复杂优化问题的求解方面有较为广泛的应用^[5,6]。该算法通过对生物种群中的个体进行遗传和变异操作,利用遗传基因算子的交叉和突变实现个体间的信息交换,经过生物群体的逐代演化,达到逼近最

优解的目的。该算法是以群体迭代为基础的,它一般以随机的原始种群作为进化起点,采用适应函数对进化结果进行筛选,根据自然界“优胜劣汰”的进化法则,利用基因算子进行选择、配对、变异等一系列遗传操作,确定次代生物群体,如此往复,直到最优解出现为止。GA算法作为一种与问题无关的求解模式,应用简单、鲁棒性强、适合进行并行处理,具有较强的搜索全局最优解的能力,可以更好地解决搜寻最优解和探寻搜索空间之间的矛盾,便于实现算法融合。然而,基本遗传算法(Simple Genetic Algorithm, SGA)的局部搜索能力较差、收敛速度慢、稳定性差且容易发生早熟收敛。针对以上存在问题,急需一种新的遗传算法来克服这种缺陷。

二叉树是一种常见数据结构,由于其二值性,可直观模拟遗传算子的二值性。

文中提出一种基于加权二叉树的自适应遗传算法,通过定义加权二叉树模型构建遗传和变异二叉树,

收稿日期:2010-03-31;修回日期:2010-06-25

基金项目:陕西省教育重点项目(09Z09)

作者简介:李龙龙(1983-),男,陕西渭南人,硕士,研究方向为智能信息系统等;王美丽,博士,研究方向为图像处理、智能系统等。

并利用蚁群信息素对不同的遗传因子进行赋权,根据权重择优选择后代进行交叉,最后采用自适应排序选择最优解,大大提高了基本遗传算法的局部搜索能力和收敛速度。

1 加权二叉树的基本定义

二叉树作为树型结构的一种特殊形式,以它在查找或检索技术中的时空有效性而得到广泛应用。文中采用加权二叉树来研究遗传算法,模型定义如下:

1.1 二叉树模型的定义

定义1 一棵二叉树是一个结点的集合 T :

(1) 如果 T 是空集,则称 T 是空二叉树;

(2) 如果 T 为有限集,则 T 可被划分为3个不相交的子集:

$\{R\}$ 根结点;

$\{L_1, L_2, \dots, L_m\}$ T 的左子树;

$\{R_1, R_2, \dots, R_n\}$ T 的右子树。

上述定义是一个递归过程,子集本身也是二叉树, T 的左子树和右子树集合可以为空,因此,立体二叉树共有5种基本形态。

定义2 子节点:如果结点 P 是结点 Q 左子树的根,则称 P 为 Q 的左子结点;如果结点 P 是结点 Q 右子树的根,则称 P 为 Q 的右子结点。

定义3 完全二叉树:对树中所有结点,用自然数进行连续编号(以根结点编号为1起,自上而下,每层以左右为序),则仅允许编号最大的非叶子结点可以没有右子结点或者没有左子结点,而小于此编号的所有结点都必须有左右2个子结点。

定义4 满二叉树:一棵深度为 h 且有 $2^h - 1$ 个结点的二叉树称为满二叉树。

1.2 蚁群信息素加权

蚂蚁之所以在寻找食物或遇到障碍物时能在最短时间找到最优路径,原因在于搜寻食物的蚂蚁能够在其所经过的搜寻路线上释放出生物信息素,作为后来蚂蚁的寻找最佳路径的信号。在群体寻径的过程中,哪条道路上积累的信息素浓度越高,说明该路径在搜寻的过程中被大多数的蚂蚁选中,其搜索性能越优,该道路被后续寻径蚂蚁选中的概率就更大,在此基础上形成一个正反馈式的机器学习过程,从而达到快速逼近最优解的目的^[7]。

引入正反馈机制虽然增强了算法的快速搜索能力,但是却会造成局部最优的结果,导致搜索停滞,为了能及时跳出局部最优,搜寻到全局最优解,对蚁群进行信息素^[8]加权就显得尤为重要。

典型的 NP 求解问题^[9] 可以直观地描述为:搜寻 m 个结点的最短遍历路径,即搜索整数结点集 $N = \{1, 2, \dots, m\}$ 的一个有向序列(即一次遍历) $A = (N_1, N_2, \dots, N_m)$,满足:

$$\sum_{i=1}^{m-1} d(N_i, N_{i+1}) + d(N_i, N_1) \quad (1)$$

取最小值。其中,结点 N_i 和 N_j 之间的距离用 $d(N_i, N_j)$ 表示。

设蚁群中共有 m 只蚂蚁,用 τ_{rs} 表示 t 时刻 e_{rs} 边上的信息素浓度, $\tau_{rs}(0) = \tau(0)$, 其中 τ_0 是常数。在完成一次遍历搜索后,需要添加新的信息素作为信息补偿,去除旧的无用的信息素,各边的信息素可调整如下:

$$\tau_{rs}(t+1) = \rho \cdot \tau_{rs}(t) + \sum_{k=1}^m \Delta\tau_{rs}^k \quad (2)$$

其中, $1 - \rho$ 为信息素的老化速度, $\Delta\tau_{rs}^k$ 为第 k 只蚂蚁在遍历边 e_{rs} 时释放的信息素补偿量。

E_k 为此次遍历过程中第 k 只蚂蚁遍历过的所有边的集合, L_k 为该蚂蚁遍历过的路径长度。蚂蚁从所处结点 r 转向移动结点 s 的概率 P_{rs}^k 可以表示为:

$$P_{rs}^k = \begin{cases} \frac{\tau_{rs}^\alpha \eta_{rs}^\beta}{\sum_{q \in A_k} \tau_{rq}^\alpha \eta_{rq}^\beta} & s \in A_k \\ 0 & \text{其它} \end{cases} \quad (3)$$

A_k 为蚂蚁 k 目前能够选择的所有结点的集合, tu_k 为蚂蚁 k 路过的所有结点集合,边 e_{rs} 的可见度用 η_{rs} 表示, α, β 为信息素及可见度在蚂蚁决策时的影响权重。

1.3 自适应排序

在对进化结果的线性排序选择中,设 N 为某个种群的规模,先按适应值对该种群降序排列;则群体中最优个体经过选择操作后的预期数量 $\eta^+ = N \times P_s(1)$;最弱个体在进行选择操作后的预期数量 $\eta^- = N \times P_s(N)$;其余个体的预期数量 $\Delta\eta = \eta_i - \eta_{i-1} = -\frac{\eta^+ - \eta^-}{N-1}$,由此可得,该算法在线性排序下的选择概率可以表示为:

$$P_s(j) = \frac{1}{N} \left[\eta^+ - \frac{\eta^+ - \eta^-}{N-1} (j-1) \right], j = 1, 2, \dots, N \quad (4)$$

其中, $1 \leq \eta^+ \leq 2$ 。若 $\eta^+ = 2, \eta^- = 0$, 则最弱个体在次代的预期数量为0,要进行群体的筛选难度最高;而如果 $\eta^+ = \eta^- = 1$, 则是按照均匀分布的情况进行随机的选择,种群的筛选难度最低^[10]。因此,为了实现种群筛选难度的动态调整,可以通过人为改变 η^+ , 使其取1和2之间不同值的方式实现,进行自适应调整的公式表示如下:

$$\eta^+ = 1 + \frac{f_{avg}}{f_{max}} \quad (5)$$

其中, f_{avg} 为每一代群体的平均适应值; f_{max} 为每一代群体中最优个体的适应值。当 $\frac{f_{avg}}{f_{max}} \rightarrow 0$ 时, $\eta^+ \rightarrow 1$; 当 $\frac{f_{avg}}{f_{max}} \rightarrow 1$ 时, $\eta^+ \rightarrow 2$, 所以, η^+ 能够随着 $\frac{f_{avg}}{f_{max}}$ 的值进行自适应调节, 进而实现动态调整种群筛选难度的目的。

由于是以随机方式产生的原始种群, 各个个体之间在进化初期的时候特性差异较大, 种群中的最优个体的适应值与平均适应值相差很多, 即 η^+ 较小, 筛选难度也较低。如果进化初期 $f_{avg} \ll f_{max}$, 则 $\eta^+ \rightarrow 1$, 选择配对方式趋近于随机方式, 筛选难度很低, 可以给较弱个体提供生存保证, 防止出现早熟收敛现象。进入进化中期, 为了对筛选难度进行动态调节, 将伴随种群进化过程中性状的动态变化而改变。到了进化后期, 种群中的平均适应值将趋近于最优适应值, 即 $f_{avg} \approx f_{max}$, 则 $\eta^+ \rightarrow 2$, 筛选难度很高, 此时对算法的求精能力有较高的要求, 从而保证算法能够快速收敛至全局最优^[11]。

2 基于加权二叉树的自适应遗传算法

文中使用二叉树构建遗传因子, 并使用蚁群信息素对其赋权, 构造加权二叉树, 并通过生成满加权二叉树的办法实现变异操作, 剪除相应的树枝结点并互换即可实现交叉操作, 后代采用自适应排序算法实现自然选择中的优胜劣汰。算法具体步骤如下:

2.1 构建遗传与变异二叉树

要构建遗传二叉树, 首先是对种群中个体染色体进行编码的问题, 目前, 浮点数编码和整数编码是两种最常见的染色体编码方式。一般来说, 聚类算法都属于多维性算法且数据量较大, 聚类样本的数量都比其聚类数量大得多, 所以, 多采用浮点数编码方式, 把各种类别的种群中心作为染色体进行编码。如对于一个拥有 4 种类别的聚类问题, 可以假设其数据集是 2 维的, 当其 4 个原始的聚类中心分别为 (1, 2), (5, 4), (8, 7), (2, 6) 时, 则其染色体编码可以描述为 (1, 2, 5, 4, 8, 7, 2, 6)。这种编码方式不但具有缩短染色体长度的优势, 而且可以提高算法速度, 在大规模数据的复杂聚类问题的求解过程中有不可取代的地位。

接着, 利用适应值计算种群中各个个体被选择的概率, 并使用轮盘赌的筛选方法选择个体。二叉树的深度代表聚类中心点的个数, 从根结点开始各层依次代表 1 号、2 号至最后一个聚类中心点; 如果选择该中心点的概率大于 0.5, 则选择左子树(数值为原码), 否则选择右子树(数值为反码), 依次类推生成遗传二叉树, 并生成相应满二叉树作为变异二叉树, 实现变异操

作; 例如, 上述 4 个聚类中心点概率分别为 0.2, 0.6, 0.8, 0.3, 那么生成的遗传及变异二叉树如图 1 所示。

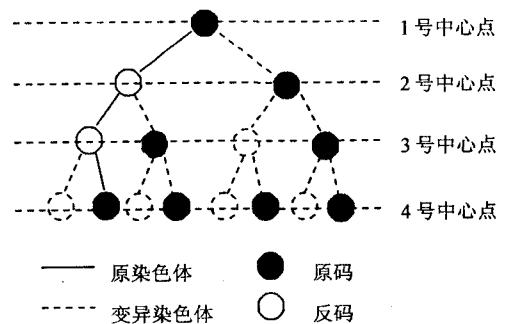


图 1 遗传与变异二叉树

2.2 二叉树赋权

生成遗传二叉树后, 接下来要利用蚁群算法获取各个路径的信息素, 即权值。蚁群算法获取信息素的过程可以描述为: 有若干个随机结点, m 只蚂蚁分别从各个结点出发, 参照转移概率 P_{ij}^k 转移到下一个结点, 已经路过的结点存入 tu_k 中。当遍历过程结束后, 以公式(2)和(3)为依据改变各条路径上的信息素浓度, 重复进行以上遍历过程, 直到信息素的变化值超出精度阈值的要求范围。

把 ACA 获得二叉树的各分支的信息素 I_m 作为二叉树各结点的权值, 与结点 N_m 共同组成加权二叉树 T , 其结点表示为:

$\text{Node}(N_m, I_m)$, m 指第 m 个结点。

成功获取权值后, 即可在原遗传与变异二叉树基础上构建加权二叉树, 如图 2 所示。

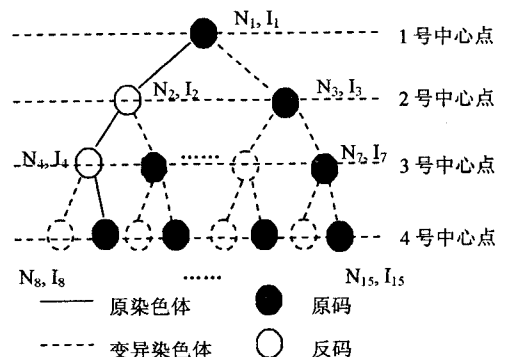


图 2 加权遗传与变异二叉树

2.3 自适应交叉操作

根据“优胜劣汰”的法则, 可以通过上一代的最优个体替换当前群体中的最弱个体的方式, 将种群中的最优解保存到子代, 从而保留适应值最高的个体, 消除适应值最低的个体。

文中采用自适应排序在遗传后代中选择优秀个体进行交叉, 并最终产生最优个体。

在交叉操作之前遍历加权二叉树生成染色体加入

交配池,在种群中任意挑选一对配子,对其进行多次交叉操作,在其生成的所有子代个体中挑选最优的两个个体,对这两个个体进行相似性判断,如果两者的相似度小于已知的相似度阈值,则将其加入子代种群;否则,对其中的次优个体进行多次局部突变操作,使二者相似度小于或等于阈值的时候,将其和最优个体放入子代种群^[12]。考虑到算法的收敛速度,加上要产生足够多有遗传价值的新个体,可使用最短距离的基因匹配算法进行算术交叉。假设两个父本个体为父本 1 和父本 2,则交配操作可以描述如下:

(1)首先比较个体父本 1 上第一个基因元素与父本 2 上各个基因元素之间的距离;

(2)找出父本 2 上与父本 1 的第一个基因元素距离最短的基因元素,并按照父本 2 个体的长度复制一条空染色体,把该基因元素放在空染色体相应的位置;

(3)使用同样的方法,逐次比较父本 1 上的其它元素和父本 2 上其余元素之间的距离,并把每次比较后距离最短的元素放在生成的空染色体上,由此即可产生父本 1 的配对个体父本 2';

(4)对父本 1 和其配对个体父本 2' 进行交叉操作:根据父本 2' 产生相应加权二叉树,在其中随机选择一个交叉结点,对其进行剪除并相互置换,即完成交叉操作。

3 实验结果与分析

文中在 Matlab 环境下编程实现检验其有效性,对 2 组现有实验测试数据进行聚类操作,其中一组数据 data1 分为 3 大类,包含 13 个样本,每个样本均有 2 个属性,分别为(0,2),(1,2),(0,1),(1,1),(6,1),(5,2),(7,3),(4,6),(5,7),(5,5),(6,5),(6,6),(5,6),数据分布如图 3 所示。

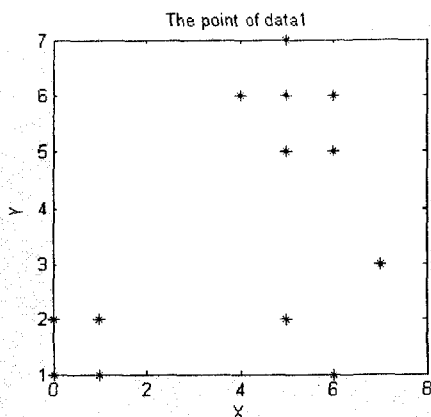


图 3 data1 数据分布图

文中的聚类效果如图 4 所示,样本 1~4 为一簇,样本 5~7 为一簇,样本 8~13 为另一簇,与普通遗传

算法得到的聚类中心是一致的。

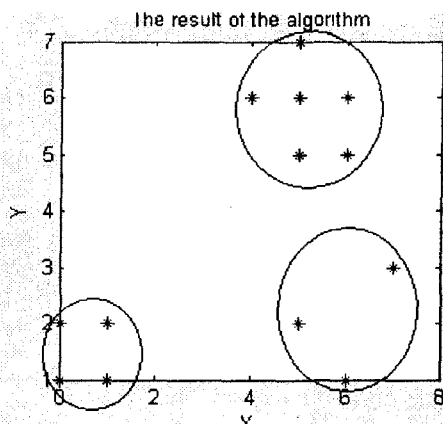


图 4 文中算法的聚类结果

第二组数据集分别为 student、teacher 以及 parents。其中,student 包含 180 个数据,分别归属于 3 个大类,每个大类分别有 60 个数据,每个数据又包含 3 个属性;数据集 teacher 拥有 260 个数据,分属于 5 大类,每个数据有 7 个属性;parents 数据集包括 245 个数据,分为 3 类,每个数据有 11 个属性,数据分布图如图 5 所示。

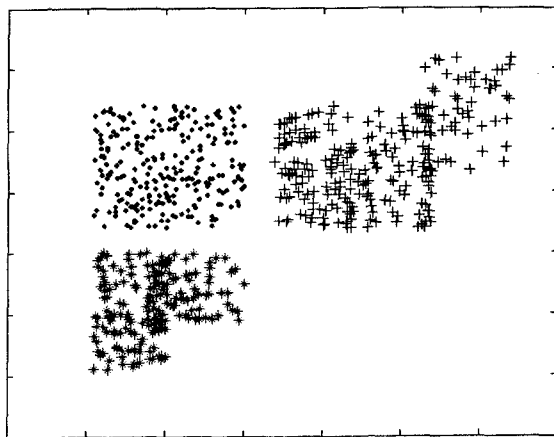


图 5 数据分布图

本算法设置种群 $m = 35$,算法的最大迭代次数 $T = 150$,所有算法运行 20 次,分别采用基本的遗传算法和文中算法进行运算,运行情况如表 1 所示。

从表 1 的结果可以看出,由于起始聚类中心点的选取对基本遗传算法的影响较大,从而导致该算法往往陷入局部最优,并因此取不到最优解,尤其是在处理较高维的数据集(如 teacher 数据集)时,没有一次达到全局最优解。然而,经过加权二叉树的算法改进后的算法在处理每一组数据集的实验中均能较为快速地收敛到最优。除数据集 student 外,基本遗传算法每组数据收敛到最优解的平均迭代次数均高于文中提出的基于加权二叉树的遗传算法,由表 1 可见,文中算法的收敛速度也比较快。

表 1 基本遗传算法与文中算法的比较

遗传算法	数据集	最大偏差	最小偏差	达到最优解的次数	达到最优解的平均迭代次数
基本遗传算法	student	79.0120	76.53114	13	4.3
	teacher	526.6243	330.35464	—	—
	parents	2.56420e5	2.198754e5	16	6.0
文中算法	student	75.6520	75.6510	14	7.8
	teacher	315.2185	315.2175	15	1.2
	parents	2.12051e5	2.12041e5	16	4.1

4 结束语

文中提出了一种基于加权二叉树的自适应遗传算法,通过构建遗传因子二叉树,利用满二叉树实现遗传因子的变异操作,并通过蚁群算法求出蚁群信息素对其进行加权,采用最短距离基因匹配的方法作为交叉操作的基础,引入自适应排序在后代中选择优秀个体进行交叉,选择最优解。实验结果表明该算法易操作、快速、准确、能够较好地克服局部最优的缺陷,大大提高了基本遗传算法的局部搜索能力和收敛速度,具有很好的应用前景。

参考文献:

- [1] Dorigo M, Maniezzo C A. The ant system: optimization by a colony of cooperating agents[J]. IEEE Trans. on System, Man and Cybernetics, 1996, 26(1): 1-13.
- [2] Dorigo M, Gambardella L M. Ant colony system: a cooperative learning approach to the traveling salesman problem[J]. IEEE Trans. on Evolutionary Computation, 1997, 1(1): 53-66.
- [3] 曹道友,程家兴. 基于改进的选择算子和交叉算子的遗传

算法[J]. 计算机技术与发展, 2010, 20(2): 44-45.

- [4] Holland J. 遗传算法的基本理论与应用[M]. 李敏强, 译. 北京: 科学出版社, 2003.
- [5] Mantawy A H, Abdel - Magid Y L, Selim S Z. Integration genetic algorithm, tabu search, and simulated annealing for the unit commitment problem[J]. IEEE Transactions on Power Systems, 1999, 14(3): 829-836.
- [6] Murthy C A, Chowdhury N. In search of optimal clusters using genetic algorithms[J]. Pattern Recognition Lett., 1996, 29(3): 825-832.
- [7] 张应辉, 王志伟, 曾庆华. 基于蚁群信息素的遗传操作算法[J]. 计算机科学, 2007, 34(6): 170-173.
- [8] 冀俊忠, 黄振, 刘椿年. 一种快速求解旅行商问题的蚁群算法[J]. 计算机研究与发展, 2009, 46(6): 968-978.
- [9] 刘芳华, 赵建民, 朱信忠. 基于改进遗传算法的物流配送路径优化的研究[J]. 计算机技术与发展, 2009, 19(7): 83-84.
- [10] 周洪伟, 原锦辉, 张来顺. 遗传算法早熟现象的改进策略[J]. 计算机工程, 2007, 33(19): 201-203.
- [11] 冯冬青, 王非, 马雁. 遗传算法中选择交叉策略的改进[J]. 计算机工程, 2008, 34(19): 189-190.
- [12] 陆林花, 王波. 一种改进的遗传聚类算法[J]. 计算机工程与应用, 2007, 43(21): 171-172.

(上接第 94 页)

- tics, 2002, 28(1): 19-36.
- [2] 周晓兰, 张杰. MATLAB 在通信系统仿真中的应用[J]. 计算机技术与发展, 2006, 16(9): 166-168.
- [3] 蔡群, 周美莲, 段杰峰, 等. 基于 Matlab 分布式工具箱的流场计算及其可视化[J]. 计算机技术与发展, 2007, 17(9): 51-54.
- [4] 哈力旦·A, 伊力哈木·亚尔买买提, 库尔班·买提木沙. 复杂背景下维吾尔文字符的分割算法[J]. 计算机工程与应用, 2007(20): 163-165.
- [5] Passoneaur, Litmand. Discourse segmentation by human and automated means[J]. Computational Linguistics, 1997, 23(1): 103-139.
- [6] 李媛, 卡米力·毛依丁. 维吾尔语笔迹鉴别方法研究[J]. 计算机技术与发展, 2008, 18(5): 9-11.
- [7] 力晓光, 李晓华, 沈兰荪. 基于纹理的图像字符自动定位技

- 术对比研究[J]. 电路与系统学报, 2006, 11(2): 258-260.
- [8] 刘晓明, 仲元红, 欧静兰. 基于 DSP 的火灾图像识别系统设计及应用[J]. 计算机技术与发展, 2006, 16(6): 96-100.
- [9] Chen Datong, Jean - Marc O, Hervé B. Text detection and recognition in images and video frames[J]. Pattern Recognition, 2004, 37(3): 595-608.
- [10] 杨述斌, 张阳. 复杂车辆图像中的车牌快速形态定位算法[J]. 计算机技术与发展, 2008, 18(6): 50-53.
- [11] 杨胜, 钟玉琢. 一种从 MPEG 压缩视频流中提取关键帧的方法[J]. 中国图像图形学报, 2001, 6(3): 254-258.
- [12] Lienhart R, Wernicke A. Localizing and segmenting text in images and videos[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2002, 12(4): 256-268.
- [13] 段汉根, 汪继文. 基于邻域滤波的图像修复[J]. 计算机技术与发展, 2007, 17(10): 34-36.