

多峰值全局优化的混合编码遗传算法

李喜艳, 周清雷, 张文宁
(郑州大学 信息工程学院, 河南 郑州 450001)

摘 要: 提出了利用黄金分割点来分割格雷码和实数编码界限的改进的混合编码遗传算法. 实验结果表明该算法能够以更快的速度找到多峰值问题的最优解或者近似最优解, 不容易陷入局部收敛, 并且具有较强跳出局部极值的能力.

关键词: 遗传算法; 黄金分割点; 格雷码; 实数编码

中图分类号: TP301

文献标识码: A

文章编号: 1000-7180(2010)03-0119-03

Mixed Coding-Based Genetic Algorithm for Multi-Peak Global Optimization

LI Xi-yan, ZHOU Qing-lei, ZHANG Wen-ning

(Information Engineering College, Zhengzhou University, Zhengzhou 450001, China)

Abstract: This paper presents an improved mixed coding-based genetic algorithm by using golden section ratio to partition Gray code and real-coded boundary. The experimental results show that this algorithm not only has higher rate to obtain optimal solution or approximate the optimal solution for multimodal problems, can not be easy to fall into local convergence, but also has stronger ability to jump out of local extremum.

Key words: genetic algorithm; golden section ratio; Gray code; real-coded

1 引言

遗传算法(GA)的卓越性能引起了人们的关注. 对于以往难以解决的函数优化问题: 复杂的目标规划问题、机器学习、图象识别、人工神经网络的权系数调整、网络构造等问题, GA 是最有效的方法之一. 虽然 GA 在许多优化问题中都有成功的应用^[1], 但其局部搜索能力差、存在未成熟收敛、对于多峰优化搜索问题进化过程中的漂移现象较常发生. 针对 GA 的不足, 很多学者提出了相应的解决方案^[1-2]. 目前, 混合遗传算法的实现方法体现在两个方面: (1) 引入局部搜索过程; (2) 增加编码变换操作过程^[3]. 在以往的遗传算法改进过程中, 很多是其他算法和 GA 的混合, 文中提出格雷码和实数编码混合编码的遗传算法, 采用 $1 \sim 0.618$ (黄金分割点) 作

为两种编码方式转换的参数, 以克服遗传算法存在的不足.

2 相关概念

程序插装方法简单地说是借助在被测程序中插入操作来实现测试目的的方法. 通常在调试程序时, 常常要在程序中插入一些打印语句, 目的就是希望执行程序时, 随时打印出关心的信息, 进而了解执行过程中程序的一些动态特征. 文中遗传算法利用程序插装技术, 可以及时得到不同的进化世代中的不同信息, 很直观的了解到了遗传算法中与实际问题紧密联系的适应度函数值的动态信息, 进而得到所要的结果.

实数编码方法指个体的每个基因值用某一范围内的一个浮点数来表示, 个体的编码长度等于其决

收稿日期: 2009-02-17; 收稿日期: 2009-05-25

基金项目: 国家“八六三”计划项目(2007AA010408)

策变量的个数.

格雷码是一种绝对编码方式,典型格雷码是一种具有反射特性和循环特性的单步自补码,其循环、单步特性消除了随机取数时出现重大误差的可能,其反射、自补特性使得求反非常方便.格雷码属于可靠性编码,是一种错误最小化的编码方式,因为该编码相邻的两个码组之间只有一位不同.

在分割时,在长度为全长的约 0.618 处进行分割,就叫做黄金分割,而分割点就叫做黄金分割点.在很多科学实验中,选取方案常用一种 0.618 法,即优选法,它可以合理地安排较少的试验次数找到合理的东西和合适的工艺条件.正是基于黄金分割点的完美分割优势,可以把不同的进化代数值映射成线段上的点,采用实数编码方法的进化代数值占有较长的线段,因此在本文中调节编码转换方式的参数 α 取值 $1 \sim 0.618^{[4]}$.

3 混合编码遗传算法

3.1 传统遗传算法的原理和步骤

遗传算法的概念最早是由 Bagley J. D 在 1967 年提出的,而开始遗传算法的理论和方法的系统性研究的是 1975 年,这一开创性工作是由 Michigan 大学的 J. H. Holland 所实行.遗传算法把问题的解表示成“染色体”,在算法中是以二进制编码的串存在.在执行遗传算法之前,给出一群“染色体”,也就是假设解.然后,把这些假设解置于问题的“环境”中,并按适者生存的原则,从中选择出适应环境的“染色体”进行复制,再通过交叉、变异过程产生更适应环境的新一代“染色体”群.这样,一代一代地进化,最后就会收敛到最适应环境的一个“染色体”上,它就是问题的最优解.

遗传算法的步骤:

- (1) 初始化.选择一个群体
- (2) 选择.以适应度为选择原则
- (3) 交叉.对于选中用于繁殖下一代的个体,随机地选择两个个体的相同位置,按交叉概率在选中的位置实行交换.
- (4) 变异.根据生物遗传中基因变异的原理,以变异概率对某些个体的某些位执行变异.

(5) 当最优个体的适应度达到给定的阈值,或者最优个体的适应度和群体适应度不再上升时,则算法的迭代过程收敛、算法结束.否则,用经过选择、交叉、变异所得到的新一代群体取代上一代群体,并返回到第(2)步即选择操作处继续循环执行.

Holland 提出的传统 GA 运用模式定理分析编码机制,建议采用操作简单的二进制编码.但是二进制编码存在着连续函数离散化时的映射误差;也不便反应所求问题的结构特征,且存在着 Hamming 悬崖问题.对于一些连续函数的优化问题,由于遗传算法的随机特性而使得局部搜索能力比较差,尤其在数值优化时难以达到精度要求,而实现时需要频繁解码,使运算过程变慢.为了克服二进制编码的缺点,对遗传算法进行改进,采用可以消除 Hamming 悬崖问题的格雷码编码方法和便于大空间搜索、可取得搜索空间任意精度的实数编码策略^[5],并且引入黄金分割点科学分割格雷码和实数编码的编码界限.

3.2 混合编码的遗传算法

文中对遗传算法的改进主要是引入黄金分割点改变基因编码方法,针对二进制编码不能直接反映问题固有的结构特征,而且二进制编码个体长度大,占用内存多,在进行数值优化时精度不高,稳定性不如表示范围较大、精度要求较高的实数编码,全局搜索能力不如格雷码编码等问题,提出了采用格雷码和实数编码混合编码的遗传算法.主要思路是首先用格雷码编码的 GA 进行较大空间的搜索,进化后期再用实数编码的 GA 进行寻优.

文中遗传算法参数主要有 5 个,即种群大小 n 、基因链长度 l 、交叉概率 p 、变异概率 q 和调节参数 α .

(1) 种群大小 n 太小时难以求出最优解,太大则增长了问题求解的收敛时间.一般 $n = 30 \sim 160$ 为宜^[6].

(2) 交叉概率 p 太小时难以向前搜索,太大则容易破坏具有高适应度值的个体结构.一般取交叉概率 $p = 0.2 \sim 0.6$ 为宜^[7-8].

(3) 变异概率 q 太小时难以产生新的基因结构,太大则使遗传算法变成了单纯的随机搜索.一般取 $q = 0.005 \sim 0.01$ 为宜^[8].

(4) 基因链长度 l 会影响遗传操作的速度,基因链越长,遗传操作越复杂,处理速度越慢;反之,基因链越短,所包含的问题参数信息越少,遗传计算速度越快.

(5) 调节参数 α 是控制格雷码和实数编码两种编码方法的转换,当进化代数小于最大代数的调节参数倍时,算法采用格雷码编码方法,并采用与格雷码编码方式对应的单点交叉、变异算子进行遗传操作,当进化代数不小于最大代数的调节参数倍时,算

法采用浮点编码方式,并采用与浮点数编码方式对应的单点交叉、变异算子进行遗传操作。

混合编码遗传算法的步骤如下:

- (1) 初始化种群;
- (2) 判断进化代数 $g < \text{最大进化代数 } mg$, 成立转向步骤(3), 不成立结束;
- (3) 判断 $g < a * mg$, 成立转向步骤(4), 不成立转向步骤(5);
- (4) 执行格雷码编码的遗传算法(二进制码转换成格雷码的法则保留二进制码的最高位为格雷码的最高位,而次高位为二进制码的高位与次高位相异或,其余各位与次高位的求法相类似,文中采用递推的方法来快速实现二进制码与格雷码的转换) $g = g + 1$, 转向(2);
- (5) 执行实数编码的遗传算法, $g = g + 1$, 转向(2).

4 实验

为了验证改进遗传算法对多峰值问题的优化效果,选择求解具有多峰值的适应度函数 $y = x * \sin(0.8\pi x)$ 的最大值,理论上混合编码的遗传算法以更快的速度寻找到的最优解或者近似最优解,不易陷入局部收敛,具有很强的跳出局部极值的能力.在实例当中选取种群规模是 50,交叉率是 0.6,变异率是 0.01, $x \in [0, 255]$. 这些数值的取值是不固定的,可以根据实际情况进行改动。

为了能够直观地了解到不同进化代数状态下适应度函数的取值情况,采用程序插装技术,在程序中插入语句:printf(“显示当前种群结果: \n”); 执行程序时我们就可以了解执行过程中程序的动态信息.表 1 就是根据这些动态信息得到的,对二进制编码的遗传算法和混合编码的遗传算法不同的世代适应度函数的最大值情况.图 1 是对应的曲线图。

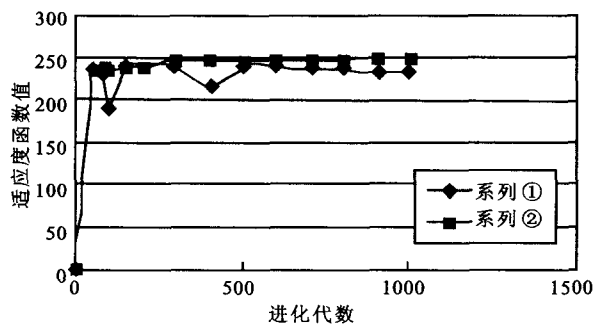


图 1 不同编码的 GA 适应度函数对比

在图 1 中,系列①表示的是二进制编码的遗传

算法适应度函数曲线,系列②表示的混合编码的遗传算法适应度函数曲线图.由表 1 结合对应的曲线,可以很直观的看出,混合编码的遗传算法在进化到 300 代的时候已经接近最优解,到了 900 代以后就逐步稳定的得到了问题的近似最优解,而二进制编码的遗传算法陷入了局部收敛,不能很好的得到问题的最优解或者近似最优解。

表 1 不同编码的 GA 适应度值的比较

进化代数	最大适应度函数值	
	二进制编码的 GA	混合编码的 GA
0	0	0
50	235.861201	235.85207698
80	231.105951	240.61213967
100	188.308671	235.85903895
150	240.616451	240.61636516
200	240.616451	240.61644111
300	240.616451	248.03615897
400	216.840197	248.03657257
500	240.616451	246.76634959
600	240.616451	248.03656160
700	235.861201	248.03636886
800	235.861201	248.03657480
900	231.105951	249.40262224
1000	231.105951	249.40269177

5 结束语

文中利用格雷码和实数编码混合编码对遗传算法进行了改进,尝试引入了黄金分割点来调整两种编码方式的转换.实验结果表明:该算法能够以更快的速度找到多峰值问题的最优解或者近似最优解,不容易陷入局部收敛,具有较强跳出局部极值的能力。

参考文献:

[1] 英海燕,王友波.基于自适应遗传算法的模板匹配研究[J].微电子学与计算机,2008,25(1):361-363.

[2] Shir O M, Back T. Dynamic niching in evolution strategies with covariance matrix adaptation[J]. Evolutionary Computation,2005,3(2/5):2584-2591.

[3] 袁健.集中新的混合遗传算法研究[D].湖南:湖南大学,2007.

[4] 李哲.多峰优化遗传算法及多目标优化进化算法研究[D].湖南:湖南大学,2007.

(下转第 125 页)

算的优点.

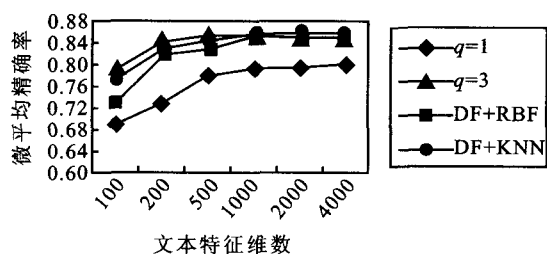


图3 微平均 F_1 坐标图

由表2可见,当特征维数高于500后,RBF分类精度并没有获得显著提升,但伴随着特征维数升高,其训练时间呈几何级数增长.可见,如果利用KPCA把特征维数压缩到500维以下,不但可以获得不逊于KNN的分类精度和好于DF的降维效果,而且可以显著降低RBF的训练时间,这正是结合KPCA和RBF神经网络的优势所在^[7].

4 结束语

特征降维是目前文本分类所面临的主要问题之一.文中提出了一种基于KPCA和RBF的文本分类新方法.实验结果表明,KPCA不仅有效实现了文本输入空间的特征降维,而且在维数低于1000的情况下取得了比DF特征选择算法更好的分类性能,同时取得了和KNN分类器相当的分类精度;如果能充分利用KPCA的这种特性,就可以在不损失分类精度的前提下显著减少RBF分类器的训练时间,从而在发挥RBF神经网络一系列优点的前提下克服其在训练集特征维数过高情况下训练时间过长的缺

点.

文中只在reuter-21578上进行了实验.在RBF神经网络训练过程中还存在过训练的缺点,这是需要改进的地方.另外,将通过对更多的语料测试集的研究来进一步明确KPCA+RBF算法在文本分类领域的能力和應用潜力.

参考文献:

- [1] 苏金树.基于机器学习的文本分类技术研究进展[J].软件学报,2006(9):1848-1859.
- [2] Yang Y, Pedersen J O. A comparative study on feature selection in text categorization[C]// Proceedings of the 14th ICML. USA: Nashville 1997:412-420.
- [3] Cover T M. The best two independent measurements are not the two best[J]. IEEE Transactions on Systems, Man and Cybernetics, 1974(4):116-117.
- [4] Scholkopf B, Smola A, Muller K R. Nonlinear component analysis as a kernel eigenvalue problem[J]. Neural Computation, 1998,10(5):1299-1319.
- [5] 崔志明,谢春丽.基于WEB的文本挖掘研究[J].微电子学与计算机,2002,19(10):51-53.
- [6] Li YH, Jain A K. Classification of the text document[J]. The Computer Journal,1998,41(8):537-546.
- [7] 王俊英,郭景峰,霍铮.中文文本分类系统的设计与实现[J].微电子学与计算机,2006,23(9):262-265.

作者简介:

杨俊男,(1983-),硕士研究生.研究方向为自然语言处理、机器学习.

陈贤富男,(1963-),博士,副教授.研究方向为机器学习、智能信息处理.

(上接第121页)

- [5] 赖鑫生,张明义.基于渗透原理迁移策略的并行遗传算法[J].计算机学报,2005,7(7):1146-1152.
- [6] 刘文强,张晓伟.多峰值全局优化的混合算法[J].计算机工程与应用,2008,44(33):43-45.
- [7] 范虹,孟庆丰,张优云.用混合编码遗传算法实现匹配追踪算法[J].西安交通大学学报,2005,39(3):295-299.
- [8] 云庆夏.进化算法[M].北京:冶金出版社,2000.

作者简介:

李喜艳女,(1983-),硕士研究生.研究方向为自动机理论和信息安全.

周清雷男,(1962-),教授,博士生导师.研究方向为自动机理论和信息安全.

张文宁女,(1982-),硕士研究生.研究方向为软件工程.