第 27 卷第 5 期　　四川理工学院学报(自然科学版)　　Vol. 27　No. 5
2014 年 10 月　　Journal of Sichuan University of Science & Engineering(Natural Science Edition)　　Oct. 2014
文章编号:1673-1549(2014)05-0043-04　　DOI:10.11863/j.suse.2014.05.11

Memetic 算法及其在分类中的应用研究

吉利鹏, 张洪伟

(成都信息工程学院计算机学院, 成都 610225)

摘　要:群体智能优化算法 Memetic 算法(Memetic Algorithm,MA)采用进化算法的操作流程,引入局部搜索算子,使其在问题的求解中保证较高收敛性能的同时又能获得较高质量的解,克服了遗传算法等传统全局优化算法易“早熟”的问题,同时避免陷入局部解。在 MA 框架基础上,提出了全局动态适应 MA 算法,采用遗传算法为全局搜索算子,k－means 算法为局部搜索算子。使用 Java 语言实现算法并对 UCI 中分类实验数据集进行测试,结果表明,将遗传算法和 k－means 结合的全局动态适应 MA 在分类问题中具有较高准确率。

关键词:Memetic Algorithm;遗传算法;局部搜索算法;分类
中图分类号:TP301.6　　**文献标志码**:A

引　言

Meme 是英国学者里查德·道金斯(Richard Dawkins)在著作《The Selfish Gene》[1]一书中首先提出,用来表示人们交流时信息传播的单元,译为“文化遗传因子”或“文化基因”。Meme 在传播过程中随个人的思想和理解而改变,因而信息在由父代传递给子代的过程中可改变,表现在算法上为局部搜索过程。

Memetic 算法(Memetic Algorithm,MA)由 Moscato 和 Norman 等[2]在 1992 年正式提出,成功应用到了 TSP 求解问题中。其作用机理和局部搜索策略在 N. Radcliffe 和 P. Surry 等撰写的论文《Formal memetic algorithms》中得到了详细的探究,此后该算法在不同的应用领域得到了广泛的研究和使用。MA 刚被提出来时,指遗传算法(Genetic Algorithm,GA)与一些局部优化算法的结合,也被称作混合遗传算法(hybrid GA)[3]。近年来,随着进化技术的不断发展,全局优化技术也得到了快速的发展,先后出现了一些基于群体的全局优化算法,如蚁群优化算法(Ant Colony Optimization,ACO)[4]和粒子群优化算法(Particle Swarm Optimization,PSO)[5],MA 算法也得到了不断的填充与丰富,并受到越来越多的国内外学者的重视[6-7]。IEEE Congress on Evolutionary Computation 等进化计算领域的权威国际会议已经把 MA 列为重要的专题进行讨论,而 2007 年国际期刊 IEEE Transactions on System, Man, and Cybernetics—Part B 也出版了关于 MA 的专刊。本文在经典 Memetic 算法的基础上,结合 k－means 局部搜索算子,提出了全局动态适应 MA 算法。

1　全局动态适应 MA 模型

全局动态适应 MA 的基本思想是:全局搜索算子采用遗传算法,局部搜索算子采用经典 k－means。利用遗传算法全局寻优的特点训练数据集,得到较优染色体,即当前较优解,使用 k－means 对较优解进行指导,反复迭代,寻找精度较高的解。每一次迭代过程,根据训练误差测试误差最小化[8],不同簇间距离最大化原则,筛选最优染色体,依据最近邻法则对测试样本进行分类预测。

收稿日期:2014-05-04
作者简介:吉利鹏(1988-),女,河南洛阳人,硕士生,主要从事计算智能方面的研究,(E-mail)ji_lipeng@163.com

2 算法的体系结构

2.1 样本归一化

设第 i 个样本为: $x_i = (x_{i1}, \cdots, x_{im})$, $x_{ij} \in R$。为了更准确地训练样本,需对所有样本进行归一化处理。方法如下:

$$x'_{ij} = \frac{[\max(\{x_{i1}, x_{i2} \ldots x_{im}\}) - x_{ij}]}{[\max(\{x_{i1}, x_{i2} \ldots x_{im}\}) - \min(\{x_{i1}, x_{i2} \ldots x_{im}\})]} \quad (1)$$

其中, m 为样本属性总数, x_{ij} 表示第 i 个样本的第 j 个属性值, x'_{ij} 表示归一化后的值。经过归一化后,所有样本的值都映射到[0,1]区间。

2.2 编码与解码

编码:输入样本集 S,总数为 N,可能被分成的簇总数为 c ,设第 i 个样本 x_i 被唯一指定在第 k_i 个簇中,即 $x_i \in C_{k_i}$, 由 $S = \bigcup_{1 \leqslant k_i \leqslant c} C_{k_i}$, 可定义染色体 e 为:

$$(x_1, x_2, \cdots, x_N) \rightarrow e$$

$$e = (k_1, k_2, \cdots, k_N), k_i \in \{1, 2, \cdots, c\} \quad (2)$$

解码:由已知染色体 e 逆推得:

$$e \rightarrow (x_1, x_2, \cdots, x_N) \quad (3)$$

2.3 适应度矢量函数

n 为样本的总数,定义距离:

$$d(x_i, x_k) = \sqrt{\sum_{j=1}^{m} (x_{ij} - x_{kj})^2} \quad (4)$$

根据簇间距离最小化和训练误差最小化原则,定义适应度矢量函数[9]:

$$F(e) = \min\{\sum_{r=1}^{c} [n_r \sum_{i=1}^{n_r} d(x_i^r, s_r)], f\}$$

$$s_r = \frac{1}{n_r}(\sum_{i=1}^{n_r} x_{i1}^r, \sum_{i=1}^{n_r} x_{i2}^r, \cdots, \sum_{i=1}^{n_r} x_{im}^r) \quad (5)$$

其中, f 为训练误差, x_i^r 表示第 r 簇中的第 i 个样本, n_r 为第 r 簇中样本总数, s_r 为第 r 簇中心, c 是簇总数。

2.4 分类算法

由最优染色体计算各簇中心 s_r, 根据最近邻法则进行分类,即按照输入的样本离簇中心 s_r 的最近距离分类[10]:

$$r_0 = \arg_r(\min\{d(x_i, s_r) = \sum_{j=1}^{m} (x_{ij} - s_{rj})^2\}) \quad (6)$$

2.5 分类器的性能评价

采用基本的分类评价指标:正确接受、正确拒绝、错误接受、错误拒绝[11]对分类预测结果进行评价。

2.6 改进 Memetic 算法流程

输入:训练集 $S = \{x_i = X_i \mid_x = (x_{i1}, \cdots, x_{im}) i = 1 \cdots N\}$。

输出:次优簇中心。

求解过程为:

Step1:随机产生染色体种群

$$Pop = \{c_p = (c_p^1, c_p^2, \cdots, c_p^K) \mid p = 1, \cdots, P, c_p^k = (c_{p1}^k, \cdots, c_{pm}^k), c_{pj}^k \in \mathbb{R}\}$$

Step2:计算 x_i 与所有簇间的距离并按最近邻原则将 x_i 划分到相应的簇,计算适应度函数 $f(c_p)$。

Step3:采用 3 - 联赛法得较优染色体种群 $Pop1 = \{c_p\}$。

Step4:对 $Pop1$ 进行交叉操作得 $Pop2 = \{c_p\}$。

Step5:对 $Pop1$ 进行变异操作得 $Pop3 = \{c_p\}$。

Step6:由 $f(c_p)$ 从 $Pop1 \cup Pop2 \cup Pop3$ 中选出较优染色体集合 $Pop4$。

Step7:对 $Pop4$ 中染色体进行 k - means 操作得 $Pop5$。

Step8:由 $f(c_p)$ 从 $Pop5$ 中选出最优染色体 c_p, 得 $Pop6$。

Step9:若结束,输出 $Pop6$, 否则用 $Pop6$ 中最优染色体替换 Pop 中最差的染色体,转到 Step2。

3 实验与结果分析

试验环境: CPU 为 Intel i5 1.70 GHz ,4 G 内存, Win7 64 位操作系统。

3.1 实验数据

选用 UCI Machine Learning Repository (http://archive. ics. uci. edu/ml/datasets) 中经典的分类数据集 Iris Dataset 和 wine Dataset 进行实验。归一化数据集,以便消除各维数据间的数量级差别,避免数据各维度之间由于数量级差别过大而造成误差过大。采用 10 - fold cross - validation 评价算法的正确性。

3.2 实验结果分析

3.2.1 Iris Dataset 实验结果

Iris 数据集中包含 150 个数据样本,三个类别为 Setosa、Versicolour 以及 Virginia,依次用 1、2、3 表示,四个样本属性为花瓣长度、花瓣宽度、萼片长度、萼片宽度分别用 PL、PW、SL、SW 表示,样本属性值均为连续型数值。每类样本中各包含 50 个数据。在三种类别中,Setosa 与其它两类没有交迭,易分离,而 Versicolour 和 Virginia 之间存在交迭,不易分离。Iris 部分数据集见表 1。

实验中算法迭代 500 次,取 Iris 中 135 样本为训练样本,剩余 15 个为测试样本。实验得最优染色体,对应的簇中心及类标签见表 2,15 个预测样本的预测结果见表 3。

表 1　Iris 数据集

PL	*PW*	*SL*	*SW*	类标签
1.4	0.2	5.1	3.5	1
1.4	0.2	4.9	3	1
5	1.7	6.7	3	2
4.5	1.5	6	2.9	2
3.5	1	5.7	2.6	2
6	2.5	6.3	3.3	3
5.9	2.1	7.1	3	3
5.6	1.8	6.3	2.9	3
5.8	2.2	6.5	3	3

表 2　最优染色体的簇中心、类标签

簇号	簇中心				类标
1	6.45	2.98	4.61	1.43	2
2	6.02	2.69	4.99	1.77	3
3	5.52	2.60	3.97	1.22	2
4	5.01	3.42	1.46	0.24	1
5	6.63	3.07	5.60	2.12	3
6	7.54	3.14	6.39	2.09	3

表 3　预测结果

簇号	预测样本编号	实际类标签	预测类标签
1	8	2	2
2	12,15	3	3
3	6,7,9,10	2	2
4	1,2,3,4,5	1	1
5	11,13,14	3	3

由表3 知,15 个预测样本均预测正确。表4 为 10 – fold cross – validation 的实验结果,计算得平均准确率为 97.33%。由表 5 知,改进 MA 算法优于其它算法。

表 4　10 – fold cross – validation

10 – 折交叉次数	预测错误样本数	准确率
1	0	100%
2	0	100%
3	0	100%
4	1	14/15
5	2	13/15
6	0	100%
7	0	100%
8	1	14/15
9	0	100%
10	0	100%

表 5　Iris 平均分类准确率比较表

Native Bayes	Boost C4.5	Bayesian Network	改进 MA
95.53%	94.33%	93.20%	97.33%

实验中采用正确接受(TP)、正确拒绝(FP)、错误接受(TN)、错误拒绝(FN)的评价指标对 MA 和 k – means 算法进行对比。结果见表 6 与表 7。

表 6　Iris 评价指标表

算法	C	TP	FP	TN	FN
k – means	1	45.0	0.0	90.0	0.0
	2	41.0	6.0	84.0	4.0
	3	39.0	4.0	86.0	6.0
改进 MA	1	45.0	0.0	90.0	0.0
	2	27.0	2.0	90.0	16.0
	3	46.0	16.0	71.0	2.0

表 7　Iris 分类结果

算法	正确分类个数	错误分类个数
k – means	13	2
MA	15	0

从表 6 可得到结论:对 Iris 数据集中三类样本分类,改进 MA 对训练集中类别 1 和类别 3 的样本全部正确分类,即准确率为 100%,而对类别 2 的分类准确率较低;k – means 仅对类别 1 的样本分类准确率为 100%。由表 7 预测结果,计算得到改进 MA 预测准确率达 100%,而 k – means 预测准确率仅 86.67%。可得结论:在分类正确率方面,改进 MA 优于 k – means。

3.2.2　Wine Dataset 实验结果

wine 数据集是通过对意大利某地产的 3 种不同品种的葡萄酒进行化学分析得到的数据,共 13 个属性,依次为 Alcohol、Malic acid、Ash、Alcalinity of ash、Magnesium、Total phenols、Flavanoids、Nonflavanoid phenols、Proanthocyanins、Color intensity、Hue、OD280/OD315 of diluted wines、Proline,类 1、2、3 依次有 59、71、48 个样本。实验结果见表 8 和表 9。

表 8　wine 评价指标表

算法	C	TP	FP	TN	FN
k – means	1	46.0	4.0	77.0	8.0
	2	47.0	5.0	64.0	19.0
	3	10.0	23.0	97.0	5.0
改进 MA	1	52.0	7.0	96.0	5.0
	2	43.0	17.0	83.0	17.0
	3	23.0	18.0	20.0	99.0

表 9　wine 分类结果

算法	正确分类个数	错误分类个数
k – means	9	6
改进 MA	14	1

从表 8 可得到结论:由 TP 指标可知,改进 MA 对训练集中类别 1 和类别 3 的分类正确率远远大于 k – means 的分类结果。由表 9 预测结果,计算得到改进 MA 预测准确率达 93.33%,而 k – means 预测准确率仅 60%。亦可得结论:在分类正确率方面,改进 MA 远优于 k – means。

4 结束语

针对进化算法“早熟”问题和局部算法易陷入局部优问题,提出将遗传算法与 k – means 算法结合的改进 MA,实验表明,该算法在收敛性和分类问题中表现较优。对 UCI 数据集的分类预测实验表明该算法较 k – means 等算法有明显优势,为数值分类预测问题提供了一定的参考。

参 考 文 献:

[1] Richard D.The Selfish Gene[M].Oxford:Oxford University Press,1976.

[2] Moscato P,Norman M G.A memetic approach for the Travelling Salesman Problem implementation of a computational ecology for combinatorial optimization on message-passing systems[C]// Valero M,Onate E, jane M, et al. Proceedings of the International Conference on Parallel Computing and Transport Applications,Amsterdam:IOS press,1992:177-186.

[3] Oh I S,Lee J S,Moon B R.Hybrid genetic algorithms for feature selection [J]. IEEE Transactions On Pattern Analysis and Machine Intelligence,2004,26(11):1424-1437.

[4] Dorigo M,Gambardella L M.Ant colony system:A cooperative learning approach to TSP[J].IEEE Transactions on Evolutionary Computation,1997,1(1):53-66.

[5] Kennedy J,Eberhart R C.Particle swarm optimization [C]//Proceedings of IEEE International Conference on Neural Networks, Perth, November 27-December 1, 1995:1942-1948.

[6] Krasnogor N,Smith J.A tutorial for competent memetic algorithms:model,taxonomy,and design issues[J]. IEEE Transactions on Evolutionary Computation,2006,10(5):474-488.

[7] Nguyen Q H,Ong Y S,Krasnogor N.A study on the design issues of memetic algorithm[C]// Proceedings of the 2007 IEEE Congress on Evolutionary Computation (CEC2007), Singapore, September 25-28, 2007: 2390-2397.

[8] 李家成,苏一丹,覃 华,等.基于遗传算法的 K 调和均值聚类算法[J].计算机技术与发展,2013,23(9):55-58.

[9] 刘 薇,刘柏嵩,王洋洋.基于改进鱼群和 K-means 的混合聚类算法[J].计算机工程与应用,2013,49(22):119-122.

[10] 袁兴梅,杨 明,杨 杨.一种面向不平衡数据的结构化 SVM 集成分类器[J].模式识别与人工智能,2013,26(3):315-320.

[11] 王 成,刘亚峰,王新成,等.分类器的分类性能评价指标[J].电子设计工程,2011,19(8):13-15.

Memetic Algorithm and its Application Research in Classification

JI Lipeng, *ZHANG Hongwei*

(College of Computer Science & Technology, Chengdu University of Information Technology, Chengdu 610225, China)

Abstract: Memetic Algorithm (MA) is one of the swarm intelligence optimization algorithms, and it has adopted the evolutionary algorithms operation process. The local search operator is used in MA to ensure the higher convergence in the solution of the problem and higher quality solutions can be obtained, then the problem that algorithm easy to “premature” of traditional global optimization algorithms, such as genetic algorithm, is overcame, at the same time, algorithm can avoid falling into local solution. The global dynamic adaptation MA is proposed based on the MA structure, the genetic algorithm is used as global search operator, k-means algorithm is used as local search operator. The algorithm is implemented by Java language and classification experiments data sets in UCI are tested, the results show that the global dynamic adaptation of MA, which combines genetic algorithm and k-means algorithm, has higher accuracy in the classification problem.

Key words: Memetic Algorithm; genetic algorithm; local search algorithm; classification