求解聚类问题的改进人工鱼群算法

王会颖[1],章义刚[2]
(1.安徽财贸职业学院 计算机系,安徽 合肥 230061;
2.合肥学院,安徽 合肥 230022)

摘 要:聚类在数据挖掘、统计学、机器学习等很多领域都有很大应用。聚类问题可以归结为一个优化问题。人工鱼群算法(AFSA)是一种新提出的新型仿生优化算法。在分析AFSA存在不足的基础上,提出一种改进人工鱼群算法,并应用于求解聚类问题。算法保持了AFSA算法简单、易实现的特点,通过改进个体鱼的行为,并引入均匀交叉算子,将人工鱼群算法和遗传算法融合,显著提高了算法运行效率和求解质量。仿真实验取得了较好的结果。

关键词:聚类;人工鱼群算法;交叉算子;优化
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2010)03-0084-04

An Improved Artificial Fish-Swarm Algorithm of Solving Clustering Analysis Problem

WANG Hui-ying[1], ZHANG Yi-gang[2]
(1. Department of Computer Science, Anhui Finance & Trade Vocational College, Hefei 230061, China;
2. Hefei University, Hefei 230022, China)

Abstract: Clustering has its roots in many areas, including data mining, statistics, and machine learning and can be regarded as an optimization problem. Artificial fish swarm algorithm (AFSA) is a novel bio-inspired optimizing method. After analyzing the disadvantages of AFSA, presents an improved artificial fish swarm optimization algorithm of solving clustering analysis problem. By improving the artificial fish's behaviors and combining artificial fish-swarm algorithm with genetic algorithm, the algorithm is as simple for implement as AFSA, but it greatly improves the ability of seeking the global excellent result and convergence property and accuracy. The simulation results show that the algorithm is more efficient.

Key words: clustering; artificial fish swarm algorithm; crossover operator; optimization

0 引 言

模仿鱼类行为方式,文献[1,2]提出了人工鱼群算法(AFSA, Artificial Fish-Swarm Algorithm),是一种基于动物自治体[3,4]的优化方法,是集群智能思想[5]的一个具体应用,它的主要特点是不需要了解问题的特殊信息,只需要对问题进行优劣的比较,通过各人工鱼个体的局部寻优行为,最终在群体中使全局最优值突现出来,有着较快的收敛速度[2]。

数据聚类是数据挖掘中的一个重要课题,在很多领域有着广泛的应用,如模式识别、图像处理和数据压缩、破产预测、交通管理、塞车状况预测等方面都有过成功的应用等。聚类分析是按照不同对象之间的差异,通过无监督学习将样本按类似性分类,把相似性大的样本归为一类,每个聚类中心起着相应类型代表的作用。

收稿日期:2009-07-09;**修回日期**:2009-10-09
基金项目:安徽省自然科学基金项目(KJ2008B021)
作者简介:王会颖(1969-),女,安徽萧县人,硕士,讲师,研究方向为智能软件、群体智能;章义刚,副教授,研究方向为群体智能。

1 聚类分析

聚类(Clustering)是数据挖掘领域最为常见的技术之一,用于发现在数据库中未知的对象类。这种对象类划分的依据是"物以类聚",即考察个体或数据对象间的相似性,将满足相似性条件的个体或数据对象划分在一组内,不满足相似性条件的个体或数据对象划分在不同的组。通过聚类过程形成的每一个组称为一个类(Cluster)[6]。

聚类分析就是从数据中寻找数据间的相似性,并依此对数据进行分类,把数据划分到不同的类中,使各类之间的离散度尽可能大,类内的离散度尽可能小。其一般数学描述为:

设模式样本集 $X = \{X_i \mid X_i = (x_{i1}, x_{i2}, \cdots, x_{id}), i = 1,2,\cdots,n\}$ 有 n 个样本,其中样本 $X_i = (x_{i1}, x_{i2}, \cdots, x_{id})$ 为 d 维向量,样本集有 k 个模式分类。

聚类问题就是要找一个划分 $C = \{C_1, C_2, \cdots, C_k\}$,满足:$X = \bigcup_{i=1}^{k} C_i, C_i \neq \varnothing, i = 1,2,\cdots,k, C_i \cap C_j = \varnothing, i,j = 1,2,\cdots,k; i \neq j$;使类内离散度之和达到最小,即(1)式。

$$F = \min \sum_{j=1}^{k} \sum_{X_i \in C_j} d(X_i, m_j) \tag{1}$$

$$m_j = \frac{1}{\mid C_j \mid} \sum_{X_i \in C_j} X_i, j = 1,2,\cdots,k \tag{2}$$

其中,m_j 为聚类中心,$d(X_i, m_j)$ 表示第 C_j 类中模式样本 X_i 到该类中心 m_j 的欧氏距离。

2 人工鱼群聚类算法

2.1 人工鱼群聚类算法中的元素

利用人工鱼群算法求解聚类问题,关键在于人工鱼个体模型的构造,构造模型中各元素的说明如下:

(1)决策变量:人工鱼个体的状态 $A = \{a_1, a_2, \cdots, a_n\}$,其中 a_i 的值为第 i 个模式样本所属的类。如 $a_1 = 2$,表示第一个模式样本属于第2类,一条人工鱼代表一个决策变量。

(2) 目标函数:人工鱼当前状态的食物浓度,即目标函数值 $Y = \sum_{j=1}^{k} \sum_{X_i \in c_j} d(X_i, m_j)$,为各样本到其聚类中心的欧氏距离之和,其中符号定义见(1)、(2)式。

求人工鱼 $A = \{a_1, a_2, \cdots, a_n\}$ 的目标函数的算法描述如下:

① 根据 $a_1, a_2, \cdots, a_n$,统计各类中有哪些模式样本。

② 根据各类中的模式样本及其属性,利用式(2),计算各聚类中心 m_j。

③ 计算各模式样本到其聚类中心的欧氏距离之和,即目标函数的值,返回食物浓度 Y。

(3) 人工鱼之间的距离:人工鱼 $A = \{a_1, a_2, \cdots, a_n\}$ 和人工鱼 $B = \{b_1, b_2, \cdots, b_n\}$ 之间的距离,$\text{distance}(A,B) = \sum_{i=1}^{n} \text{sign}(\mid a_i - b_i \mid)$,其中:$\text{sign}(x) = \begin{cases} 0, x = 0 \\ 1, x \neq 0 \end{cases}$,表示相应不同分量的个数。

例如:设模式样本集有6个样本,$n = 6$;分为3类,$k = 3$;人工鱼 A 的状态 $A = \{1,1,2,2,3,3\}$,人工鱼 B 的状态 $B = \{1,2,3,1,2,3\}$,则 $\text{distance}(A,B) = 4$。

(4) 人工鱼的 r - 距离邻域:设人工鱼的集合为 G,则人工鱼 A 的 r - 距离邻域表示为 $N(A,r) = \{B \mid \text{distance}(A,B) <= r, B \in G\}$。

(5) 人工鱼群中心:人工鱼群 $A_1, A_2, \cdots, A_m$ 中心位置 $\text{Center}(A_1, A_2, \cdots, A_m)$ 定义如下:

$$\text{Center}(A_1, A_2, \cdots, A_m) = \underset{i=1,\cdots,m}{\text{Most}} (a_i^1, a_i^2, \cdots, a_i^n)$$

2.2 人工鱼群算法

采用面向对象的方法描述人工鱼群算法。设模式样本集中有 n 个样本,有 m 条人工鱼,人工鱼的属性有:人工鱼的视野范围 visual,拥挤因子 δ(delta),每次试探的最大次数 trynumber;人工鱼行为详细描述如下[2]:

(1) 觅食行为:设人工鱼当前状态为 X_i,在其视野范围内随机选择一个状态 X_j,如果 $Y_i < Y_j$,则向该方向前进一步;反之,再重新随机选择状态 X_j,判断是否满足前进条件;试探 trynumber 次后,如果仍不满足前进条件,则执行其他行为(如随机移动行为)。

(2) 聚群行为:设人工鱼当前状态为 X_i,探索其邻域的伙伴数目 n_f,如果 $n_f/N < \delta, (0 < \delta < 1)$,则表明伙伴中心有较多的食物并且不太拥挤,如果此时 $Y_i < Y_c$,则人工鱼向中心位置 X_c 前进一步;否则执行其他行为(如觅食行为)。

(3) 追尾行为:设人工鱼当前状态为 X_i,探索其邻域内状态最优的邻居 X_{max},如果 $Y_i < Y_{max}$,并且 X_{max} 的邻域内伙伴的数目 n_f 满足 $n_f/N < \delta, (0 < \delta < 1)$,表明 X_{max} 的附近有较多的食物并且不太拥挤,则向 X_{max} 的位置前进一步;否则执行觅食行为。

(4)公告板:公告板用来记录最优人工鱼个体的状态。各人工鱼个体在寻优过程中,每次行动完毕就检验自身状态与公告板的状态,如果自身状态优于公告板状态,就将公告板的状态改写为自身状态,这样就使公告板记录下历史最优的状态。

2.3 人工鱼群算法的改进

1)觅食行为的改进。

在2.2节的 AFSA 中,觅食行为执行的是一种随机试探行为,在试探过程中,满足前进条件,前进一步,其目的是寻找较好的解,是进步。在试探 trynumber 次后,如果仍不满足前进条件,则执行其他行为(如随机移动行为)。但是,随机移动行为很可能使解有很大的倒退,影响进步的大方向。可以进一步改进为:保存试探过程中的较优解,在试探 trynumber 次后,如果仍不满足前进条件,则向这个较优解前进一步。这样,既有利于寻优活动的全面展开,又保持了进步的大方向,易于冲出局部最优解,向全局最优解转化。

2)群聚行为和追尾行为的改进。

在求解聚类问题的人工鱼群算法中,觅食行为在每次试探后都要计算目标函数的值,计算量较大,计算

时间长。因此,群聚行为可以改进为:当人工鱼向中心位置 X_c 前进一步失败后,不再执行觅食行为,而是返回原来的值和解。追尾行为同样改进为:当向 X_{max} 的位置前进一步失败后,不再执行觅食行为,而是返回原来的值和解。这样减少了人工鱼的搜索时间,提高了算法的效率。

3)公告牌的应用。

人工鱼群算法整体表现出快速向极值域收敛的特性,其主要原因是有较好解的引导。人工鱼的群聚和追尾行为本质是追逐较好解。有了较好解的引导,人工鱼才能较好地快速前进。因此,增加公告牌的功能,除记录最优人工鱼个体的状态外,还能使之起到较好解的引导作用。在一代循环中,若所有人工鱼的行为的结果都没有改写公告牌,则随机选择一条鱼,把公告牌上的解赋给该鱼,将当前最优解带入下一代循环,起到引导作用。

4)引入动态因子和均匀交叉算子。

人工鱼群算法在应用中也存在不足。主要表现在以下两点:(1)当寻优的域较大或处于变化平坦的区域时,收敛于全局的最优解速度减慢、搜索性能劣化;(2)算法一般在优化初期具有较快的收敛性,后期却往往收敛较慢[7]。针对这些缺陷,笔者引入动态因子和均匀交叉算子来改进人工鱼群算法的不足。

遗传算法中的均匀交叉过程是[8]:先随机地产生一个与父辈个体基因串具有同样长度的二进制串,该串中 0 表示不交换,1 表示交换。该二进制串称为交叉模板,然后根据该模板对两个父辈基因串进行交叉,得到两个新基因串,即为后代新个体。例如:

父辈个体 A: 10110111001　　父辈个体 B: 00101100100

交叉模板: 10011011100

→新个体 A': 00101100101　　新个体 B': 10110111000

例如:假如有 10 个模式样本,需要聚成 3 类,公告牌上记录的最优人工鱼个体的状态 A - best 为:1123222333,其中 1,2,3 表示样本所属的类,位置表示第几个样本,即样本 1 聚在第 1 类,样本 2 聚在第 1 类,样本 3 聚在第 2 类,样本 4 聚在第 3 类等等。人工鱼个体 i 的状态 A_i 为:2131222333,均匀交叉过程如下:

A - best: 1123222333　　A_i:2131222333

交叉模板: 1001101110

→ 新个体 A': 2121222333　　新个体 B': 11332223335

人工鱼群算法中的均匀交叉过程:

① 公告牌上记录最优人工鱼个体的状态 A - best 和第 i 条人工鱼的当前状态 A_i 作为两个父辈基因串,进行均匀交叉,得到两个新基因串个体 A' 和 B';

② 计算这两个新个体 A' 和 B' 的目标函数值,若优于 A - best 的目标函数值,则改写公告牌,记录下当前最优状态,结束交叉过程,否则转 ①,直到重复一定的次数;

③ 若 m 条人工鱼都执行了①、②两步,则结束交叉过程,否则转①。

针对人工鱼群算法在优化初期收敛速度较快,后期收敛较慢的缺陷,在算法中引入动态因子。动态因子记录在进化过程中公告牌连续没有发生变化的代数。当后期收敛速度缓慢到一定程度,即动态因子达到一定的阈值时,动态地激活人工鱼群算法中的均匀交叉过程,对算法进行优化,改善算法的收敛速度。均匀交叉算子容易产生很多随机的解,且对原有解有很大的突破,兼顾解空间的多种情况,有效地改善了算法的速度减慢、搜索性能劣化的不足。

5)移动策略的选择。

在求解聚类问题的人工鱼群算法中,因为相应元素计算量较大,移动策略按照有进步即可的原则。移动策略为:先进行追尾行为,若没有改变公告牌,则执行群聚行为,若再没有改变公告牌,则执行觅食行为。

2.4 整体算法描述

基于上述对人工鱼群算法的改进,提出求解聚类问题的改进人工鱼群算法(IASFA),算法的整体描述如下:

① 初始化。设定人工鱼的数量 m,视野范围 visual,拥挤因子 δ(delta),每次试探的最大次数 trynumber,动态因子的阈值 q_0,最大进化代数 maxgen 等参数;初始化 m 条鱼的初始状态 A[m],计算其目标函数值 Y[m],登记公告牌。

② 每条鱼先进行追尾行为,若状态优于公告牌,则更新公告牌后转 ③,否则执行群聚行为;若状态优于公告牌,则更新公告牌后转 ③,否则,执行觅食行为;若状态优于公告牌,则更新公告牌。

③ 若 m 条鱼都执行了第 ② 步,则转 ④,否则转 ②。

④ 若公告牌没有更新,动态因子加 1,若动态因子达到其阈值 q_0,则执行人工鱼群算法中的均匀交叉过程,否则,随机选择一条鱼使其状态改变为公告牌上的状态,将当前最优解带入下一代循环,实现公告牌的引导作用。

⑤ 若这一代循环中,公告牌进行了更新,则动态因子清零,进化代数 gen = gen + 1,若 gen < maxgen,则

转②,否则结束本算法。

3　仿真实验

实验数据来源于 UCI 机器学习数据库中的 3 个经典数据集:iris, wine, glass 数据集。实验运行环境为:Pentium 4, 2.26G CPU, 512 内存,Eclipse－SDK－3.4.1,Java 编程。实验参数:visual＝100,delta＝0.75,trynumber＝300,maxgen＝600,鱼数 $m = 20, q_0 = 10$。

实验采用 k 均值算法(k－means),人工鱼群算法(AFSA),改进后的人工鱼群算法(IAFSA)对上述 3 个数据集进行聚类分析,实验结果如表 1,表 2,表 3 和图 1,图 2 所示。

表 1　算法 AFSA 和 IAFSA 求解结果的比较

Iris				Wine				Glass			
AFSA		IAFSA		AFSA		IAFSA		AFSA		IAFSA	
值	代数	值	代数	值	代数	值	代数	值	代数	值	代数
97.43	204	97.22	221	16978.26	252	16530.53	196	220.67	253	217.17	156
97.51	300	97.22	178	16776.61	216	16530.53	192	225.89	213	213.20	155
97.23	256	97.22	165	16578.76	240	16530.53	191	220.26	281	213.20	155
129.2	279	97.22	178	16530.53	230	16530.53	169	251.68	273	213.20	151
133.7	204	97.22	211	17069.15	296	16530.53	160	225.88	204	213.22	150
97.74	139	97.22	135	16985.83	261	16530.53	142	219.29	210	215.21	146
135.51	449	97.22	251	17222.82	211	16530.53	167	222.14	253	213.20	159
97.73	246	97.22	215	16952.99	193	16530.53	151	218.23	247	213.22	161
97.22	169	97.22	197	16536.19	261	16530.53	170	213.22	243	213.20	153
126.9	276	97.22	131	16548.91	297	16530.53	191	222.31	229	215.98	159

表 2　算法的统计结果比较

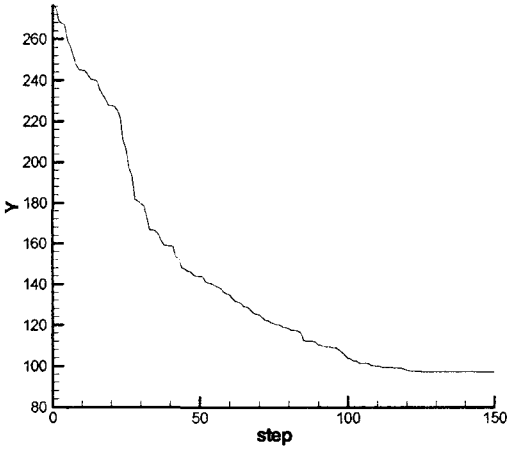
算法	数据	最优值	平均值	最差值	最优解次数	数据集提供的分类结果求得的目标函数值
K－means	Iris	97.22	102.89	124.18	2	Iris:100.51
	Wine	16530.53	16913.71	18436.95	1	
	Glass	213.22	224.82	257.83	0	
AFSA	Iris	97.22	115.22	142.00	14	Wine:23969.07
	Wine	16530.53	17007.50	20827.50	3	
	Glass	213.22	223.13	257.19	0	
IAFSA	Iris	97.22	101.07	121.27	100	Glass:334.78
	Wine	16530.53	16531.16	16587.98	98	
	Glass	213.20	216.670	252.11	67	

表 3　均匀交叉前后值的比较

数据	交叉前	交叉后	改进值	数据	交叉前	交叉后	改进值
Iris	108.89	97.99	10.90	Iris	103.30	98.15	5.15
	107.95	99.24	8.71		102.02	97.329	4.691
	104.26	97.51	6.75		99.56	97.32	2.24
Wine	18260.56	17702.02	558.54	Wine	18572.85	17597.82	975.03
	18069.11	17453.56	615.55		18322.56	16651.38	1671.18
	18922.29	18107.65	814.64		17208.08	16668.37	539.71
Glass	231.31	224.40	6.91	Glass	229.43	219.40	10.03
	230.81	224.41	6.40		227.40	219.83	7.57
	230.96	222.77	8.19		221.27	218.98	2.29

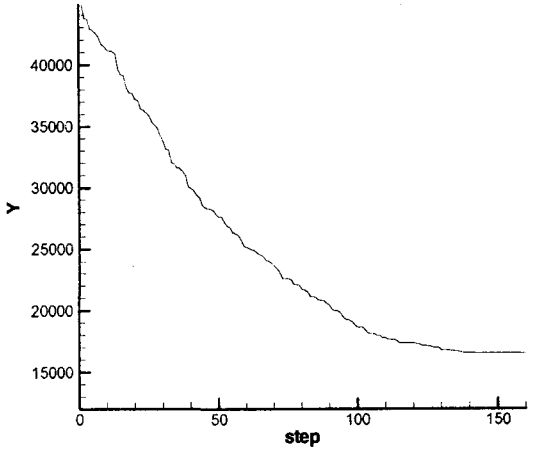
表 1 为算法 AFSA 和 IAFSA 连续 10 次的运算结果,表 2 为算法 AFSA 和 IAFSA 连续 100 次的运算结果的统计。表 2 中 IAFSA 求解的最优值、平均值、最差值等各指标均优于 k－均值和 AFSA,优于原数据集提供的分类结果求得的目标函数值。可以看出 IAFSA 能容易收敛到最优解,求解结果稳定,通用性较强。表 2 中,最优解次数为连续 100 次求解中各算法收敛到最优解的次数,对 iris、wine 数据集,IAFSA 收敛到最优解的次数达到 100,98,67 次,而 k－均值和 AFSA 仅达到几次,差别明显,说明 IAFSA 容易突破局部最优解,收敛到全局最优解。表 3 反映了均匀交叉算子对解的改进情况,体现了均匀交叉的有效性。

图 1,2,分别以 iris、wine 数据集为例,用 IAFSA 求解,收敛到最优解时,解的进化情况。横坐标为进化代数 step,纵坐标为目标函数值 Y。进化初期收敛较快,后期变化充分,最后收敛到最优解。



(131 代以后目标函数值稳定在 97.22)

图 1　Iris 数据 IAFSA 聚类过程



(142 代以后目标函数值稳定在 16530.53)

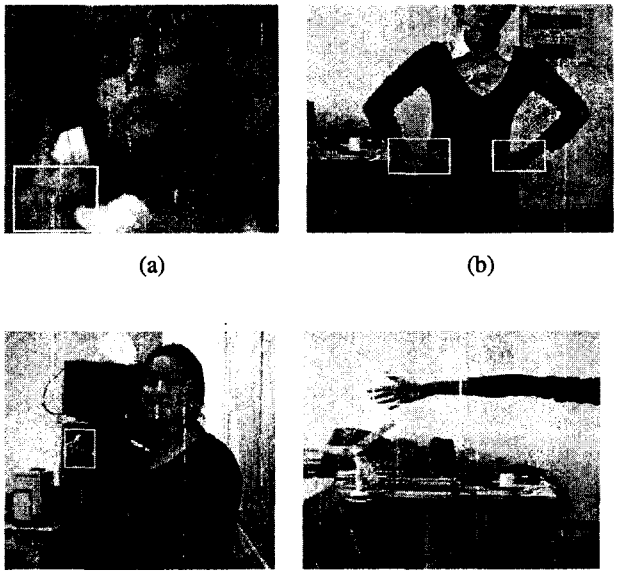
图 2　Wine 数据 IAFSA 聚类过程

4　结束语

文中研究在分类数目已知,利用目标函数方法的聚类分析。在聚类算法上,采用人工鱼群算法。对其

(下转第 91 页)

以后将逐步完善该检测算法,利用人脸和人手的纹理特征加以区别,提高跟踪准确率。



(a)　(b)　(c)　(d)

图4　部分实验结果

参考文献:

[1] Wang Liang, Hu Weiming, Tan Tieniu. Recent Developments in human motion analysis[J]. Pattern recognition, 2003, 36(33):585-601.

[2] 路　凯,李小坚,周金祥.基于肤色和边缘轮廓检测的手势识别[J].北方工业大学学报,2006,18(3):12-15.

[3] 江冬梅,吴晓娟.基于肤色和特征相似度映射手势识别算法[J].电子测量技术,2004(2):27-29.

[4] Hani A, Almohair K, Ramli A R, et al. Skin Detection in Luminance Images using Threshold Technique[J]. International Journal of the Computer, the Internet and Management, 2007,15(1):25-32.

[5] 齐苏敏,黄贤武,刘家盛.利用基于颜色的自适应形状模型实现手势跟踪[J].计算机应用研究,2008,25(2):485-488.

[6] Brand J, Mason J S. A Comparative Assessment of Three Approaches to Pixel-level Human Skin-Detection[C]//Proceedings of 15th International Conference on Pattern Recognition. Barcelona:[s.n.],2000:1056-1059.

[7] Brand J D, Mason J S D. Skin Probability Map and its use in Face Detection[C]//Proceedings of 2001 International Conference on Image Processing. Thessaloniki, Greece:[s.n.], 2001:1034-1037.

[8] Zhang Ming ji, Wang Wei qiang, Zheng Qing fang, et al. Skin-Color Detection Based on Adaptive Thresholds[C]//Proceedings of the Third International Conference on Image and Graphics. Hong Kong:[s.n.],2004:250-253.

[9] 王　娜,杜世培.彩色地图种道路的识别和抽取[J].计算机工程与设计,2007,28:2642-2645.

[10] 周　桢.复杂场景中目标抗遮挡跟踪算法研究[J].航空兵器,2007(6):72-75.

[11] 孔晓明,陈一民,陈养彬,等.基于视觉的动态识别[J].计算机工程与设计,2005,26:2934-2936.

(上接第87页)

进行改进,提出了求解聚类问题的改进人工鱼群算法。算法的改进体现在:觅食行为、群聚行为、追尾行为的改进;公告牌的应用;将遗传算法和人工鱼群算法融合,引入动态因子和均匀交叉算子。觅食行为的改进,既有利于寻优活动的全面展开,又保持了进步的大方向。群聚行为、追尾行为的改进,减少了人工鱼的搜索时间,提高算法了的效率。公告牌的应用,使当前最优解能带入下一代循环,起到引导作用。动态因子和均匀交叉算子的引入有效改进了算法在优化初期具有较快的收敛性,后期却往往收敛较慢的缺陷。实验选用UCI机器学习数据库中的iris, wine, glass经典数据集,对k均值算法(k-means),人工鱼群算法(AFSA),改进的人工鱼群算法(IAFSA)进行测试。实验表明IAFSA产生了上述作用,求解结果稳定,通用性较强,取得较好的效果。进一步需要解决的问题是:研究人工鱼群算法中各参数对算法的影响,增强算法稳定性、通用性,提高算法的求解精度和收敛速度。

参考文献:

[1] 李晓磊,邵之江,钱积新.一种基于动物自治体的寻优模式:鱼群算法[J].系统工程理论与实践,2002,22(11):32-38.

[2] 李晓磊,路　飞,田国会,等.组合优化问题的人工鱼群算法应用[J].山东大学学报:工学版,2004,34(5):64-67.

[3] Wilsons. The animal path to AI[C]//Proceedings of the First International Conference on the Simulation of Adaptive Behavior. Cambridge: MIT Press,1991.

[4] Jeffrey D. Animals and what they can tell us[J]. Trends in Cognitive Sciences,1998,2(2):60-67.

[5] Bonabeau E, Theraulaz G. Swarm smarts[J]. Scientific American,2000,282(3):72-79.

[6] 谢维信.工程模糊数学方法[M].西安:西安电子科技大学出版社,1991:136-165.

[7] 张梅凤,邵　诚,甘　勇,等.基于变异算子与模拟退火混合的人工鱼群优化算法[J].电子学报,2006,34(8):1381-1385.

[8] 杨善林,倪志伟.机器学习与智能决策支持系统[M].北京:科学出版社,2004:160-197.