

并行小生境粒子群优化的模糊聚类算法

常天庆, 李 勇, 杨国振, 张 雷

(装甲兵工程学院 控制工程系, 北京 100072)

摘 要: 针对模糊聚类算法对初始聚类中心敏感、容易陷入局部最优的问题, 采用并行小生境粒子群优化算法对模糊聚类算法进行改进. 通过山谷函数对小生境进行识别以形成互斥的多个子群, 采用惩罚函数实现多子群并行搜索过程中的信息共享机制, 引入混合聚类有效性函数获取最佳聚类数. 仿真结果表明, 该算法能提高模糊聚类算法的搜索效率以及分类精度.

关键词: 小生境; 粒子群; 模糊聚类; 小生境识别; 信息共享机制; 混合聚类有效性函数

中图分类号: TP301.6

文献标识码: A

文章编号: 1000-7180(2014)01-0013-04

Fuzzy Clustering Algorithm Based on Parallel Niches Particle Swarm Optimization

CHANG Tian-qing, LI Yong, YANG Guo-zhen, ZHANG Lei

(Department of Control Engineering, Academy of Armored Forces Engineering, Beijing 100072, China)

Abstract: Because fuzzy clustering algorithm is sensitive to initial cluster center and easy to fall into local optimum, improved fuzzy clustering algorithm based on parallel niches particle swarm optimization algorithm is presented in this paper. Valley function is used to form mutually exclusive subswarms by identifying niches. Punishment function is adopted to realize information sharing mechanism in the process of subswarms parallel searching. Hybrid cluster validity function is imported to obtain optimal cluster number. Simulation results show that the improved algorithm can enhance searching efficiency and classification accuracy of fuzzy clustering algorithm.

Key words: niches; particle swarm; fuzzy clustering; niches identification; information sharing mechanism; hybrid cluster validity function

1 引言

模糊聚类是用模糊集的方法定量确定样本的亲疏关系, 从而定量地确定被研究对象之间的亲疏关系, 达到对样本合理分类的目的^[1]. 模糊 C-均值聚类算法(Fuzzy C-Means Clustering Algorithm, FCMCA)是目前应用较广的模糊聚类算法. FCMCA是一种无监督的自学习过程, 实现过程相对简单而且运算速度快, 目标函数采用的是欧式距离, 具有较直观的几何含义. 但由于模糊聚类目标函数是一个非凸函数, 存在大量的局部极值点, 而 FCMCA 实际上是一种基于梯度下降的局部搜索算法, 虽然每步迭代都朝着好的方向进行, 具有很强的局部搜索

能力, 但它对初始聚类中心敏感, 容易陷入局部最优而不能得到全局最优解^[2].

针对该问题, 一些具有全局优化能力的算法被引入到 FCMCA 中形成混合模糊聚类算法, 这些算法虽然能以较高的概率收敛到全局最优点, 但均不同程度存在早熟现象. 本文将并行小生境技术引入到粒子群优化算法中形成并行小生境粒子群优化算法.

2 并行小生境粒子群优化的 FCMCA 算法

本文在串行小生境 PSO 技术的基础上, 提出了一种并行小生境粒子群优化算法(Parallel Niches PSO, PNPSO). 该算法通过山谷函数对小生境进行

识别以形成互斥的多个子群,当两个不同子群中最优粒子均位于同一个小生境时,通过比较它们之间的适应度值大小,使较大值粒子所在的子群占用该小生境,而较小值粒子所在的子群作为失败者将被重新初始化以搜索新的区域.子群中的粒子在搜索过程中,其最优位置会发生变化,如果位置的改变使之侵入到其他子群中,此时通过信息共享机制对其进行惩罚,改变其适应值大小,使之在子群中失去影响力,以维持种群多样性.

2.1 小生境识别

小生境技术需要解决的问题之一即小生境的划分,通过小生境划分将解空间分为不同的搜索域,每一个搜索域包含一个局部最优解.小生境识别则是通过一定的方法将不同的搜索域彼此分开,避免了不同的粒子群搜索同一个解空间.目前绝大部分小生境技术都是依靠一些先验知识来设定一个统一的小生境半径,从而将整个解空间划分为不同的搜索子空间.但即使通过先验知识预先设置小生境半径,对求解复杂的多模优化问题来说,算法也并不能获得好的效果.

本文采用 Ursem^[3]定义的山谷函数,作为小生境识别的依据.通常,山谷函数通过在当前运行子群中的粒子与其他子群的最优粒子之间插入多个采样点,判断两点是否属于同一个小生境.如果任何一个内部的采样点对应的适应度小于两点间的最小值,则认为这两点属于不同的小生境,反之,则认为这两点属于同一个小生境.

设 d_p 和 d_q 为搜索空间中任意两点, d_{in} 为 d_p 和 d_q 之间内部采样点,按照下式计算内部采样点:

$$d_{in} = d_p + (d_q - d_p) \cdot \text{samples}[i] \quad (1)$$

式中, samples 为采样入口数组, i 为采样数组的第 i 个入口.为保证两点间的采样密度, samples 数组长度与 d_p 和 d_q 之间的距离 $\text{dis}(d_p, d_q)$ 成正比^[4],即:

$$\text{sample length} = \text{coff} \cdot \text{dis}(d_p, d_q) \quad (2)$$

式中, coff 为比例系数.

2.2 信息共享机制

为了避免正在运行的多个粒子群重复搜索已经发现最优解的小生境,可以采用粒子群之间的信息共享机制,即不同子群之间通过共享信息,一个正在执行搜索任务的子群将不再搜索已经属于其他子群的小生境.

采用惩罚函数实现信息共享机制,即当一个子群拥有某个小生境时,其他子群侵入该小生境的个

体将受到惩罚,并使其对应的适应度值减小,这样就可以迫使侵入者搜索其他小生境.算法中使用加性死罚函数,使侵入到与已经发现的最优解位于同一个小生境的粒子在其子群中失去影响力.具体函数如下:

$$E(x_r^i) = \begin{cases} f(x_r^i) & x_r^i \in \{D(x_k) \mid k = 1, 2, \dots, r-1\} \\ f(x_r^i) - p(x_r^i) & \text{其他} \end{cases} \quad (3)$$

式中, x_r^i 为正在运行的子群中第 i 个粒子位置, r 为子群序号. $f(x_r^i)$ 为该粒子原始的适应度值, x_k 为目前已收敛的子群中最优粒子的位置, $D(x_k)$ 为 x_k 所在的小生境, k 的取值范围为 $1 \sim r-1$, $p(\cdot)$ 为罚函数,可设置为一个较大值,使改变的粒子适应度低于所有的局部最小值.

在 PSO 算法中,影响粒子移动的两个因素是个体粒子历史最优位置和种群中所有粒子的历史最优位置,通过改变这两个因素就能改变粒子的运动轨迹.引入小生境算法后,采用信息共享机制,通过罚函数改变粒子的适应度值及其位置和速度参数,从而改变粒子在种群中的影响力.

2.3 混合聚类有效性函数

传统的 FCMCA 是在聚类数 c 给定的情况下进行的.通常,聚类数是根据领域知识人为地指定,这样得出的 c 值具有很大的随意性,直接影响到聚类效果.本文引入聚类有效性评价函数,通过自适应地调整聚类数 c 的取值,使聚类有效性函数达到极值,从而获得最佳聚类数,避免在聚类数目选取上存在的主观性.

混合有效性函数^[5] v_{hybrid} 结合了 Amine M. Bensaid^[6]有效性函数 v_{bsaid} 和 Kwon^[7]有效性函数 v_{kwn} 各自的优点,对 Xie-Beni^[8]有效性函数 v_{XB} 进行改进,其有效性函数如下:

$$v_{\text{hybrid}}(U, V, c) = \frac{\sum_{i=1}^c \sum_{j=1}^n u_{ij}^m \|v_i - x_j\|^2 + \frac{c}{n^2} \sum_{k=1}^{n-1} \sum_{j=k+1}^n \|x_k - x_j\|^2}{\sum_{i=1}^c \|v_i - v_j\|^2} \quad (4)$$

在分子中引入类似于 v_{kwn} 惩罚因子 $\frac{c}{n^2} \sum_{k=1}^{n-1} \sum_{j=k+1}^n \|x_k - x_j\|^2$,而分母上使用 v_{bsaid} 中分母来代替 v_{XB} 中分母. v_{hybrid} 引入惩罚因子后避免了当 $c \rightarrow n$ 时 v_{hybrid} 丧失鲁棒性和寻找最佳聚类数目的能力,增加

了稳定性.

2.4 算法描述

并行小生境粒子群优化的 FCMCA (PNPSO-FCMCA) 的算法描述如下: 根据聚类数 c , 从样本中随机选取个样本作为聚类中心, 生成一个粒子, 同理产生拥有 N 个粒子的种群, 通过小生境识别将种群划分为互异的 m 个子粒子群, 每个子群按照 PSO 算法搜索各自小生境中的最优值. 搜索过程中实施信息共享机制, 调整各子群中粒子参数, 确保算法能够搜索到全局最优值. 同时采用 φ_{hybrid} 作为聚类有效性的评价函数, 对于给定的模糊加权指数 m , φ_{hybrid} 随着 c 的增加而减小, 当 c 达到最优值 c^* 时 φ_{hybrid} 取得最小值, 继续增大 φ_{hybrid} 反而增大, 因此当 φ_{hybrid} 取得最小值时, 对应的聚类数为最佳聚类数, 此时聚类效果最佳.

算法的具体步骤描述如下.

步骤 1: 设定聚类数 $c = 2$;

步骤 2: 生成粒子群. 从 n 个数据样本中随机选取 c 个作为聚类中心, 构成一个粒子 $V_i = (v_{i1}, v_{i2}, \dots, v_{ic})$, 同理产生拥有 N 个粒子 $V_1, V_2, \dots, V_N$ 的种群;

步骤 3: 初始化小生境粒子种群. 将 N 个粒子的种群分成 m 个子群, 初始化各子群中粒子初始位置、速度, 子群中粒子历史最优位置, 各子群中参数 c_1, c_2, ω , 迭代次数 N_c ;

步骤 4: 各子群根据山谷函数进行小生境识别, 将处于相同小生境的子群中的最优位置进行竞争, 胜利者占用小生境, 失败者调整子群中粒子参数, 搜索新的区域;

步骤 5: 计算隶属度矩阵;

步骤 6: 根据适应度函数, 求出粒子最优位置和子群最优位置, 并按照 PSO 算法更新粒子速度及位置;

步骤 7: 按照信息共享机制, 对侵入其他子群的粒子实施惩罚, 调整受惩罚粒子参数;

步骤 8: 检查终止条件(最优值不再变化或达到最大迭代次数). 如果不满足终止条件, 则转入步骤 5, 否则, 输出目前为止的全局最优解, 即最优聚类中心簇并转入步骤 9;

步骤 9: 求取混合聚类有效性函数 φ_{hybrid} ;

步骤 10: 判断 φ_{hybrid} 是否最小, 如果最小, 则转入步骤 11, 否则, $c = c + 1$ 后转入步骤 2;

步骤 11: 返回最优聚类数 c^* 以及最佳聚类中心, 根据隶属度矩阵, 将样本分类到各个最佳聚类

中心.

3 仿真实验及分析

为了验证本文提出模糊聚类算法的有效性, 采用机器学习数据库中的 IRIS 数据集进行仿真实验. 该数据集由 4 维空间的 150 个样本组成, 样本的类别数为 3.

设置初始类别数为 2, 然后逐步调整类别数, 以混合聚类有效性函数作为评价指标来确定最佳分类. 得出类别为 2 至 7 的有效性函数值如表 1 所示.

表 1 IRIS 数据集的聚类有效性函数值

聚类数	有效性评价函数 φ_{hybrid}
2	0.108 5
3	0.101 8
4	0.128 3
5	0.136 1
6	0.152 9
7	0.146 6

对 IRIS 数据集进行聚类, 在类别数为 3 时, 聚类有效性函数取得最小值, 这与实际数据集的类别数是一致的.

利用模糊聚类算法(FCMCA)、粒子群优化的模糊聚类算法(PSO-FCMCA)以及并行小生境粒子群优化的模糊聚类算法(PNPSO-FCMCA)对 IRIS 数据集进行分类, 结果分别见表 2、表 3、表 4 和表 5.

表 2 FCMCA 算法分类结果

第 i 类	分到第 i 类		
	1	2	3
1	50		
2		44	6
3		9	41

表 3 PSO-FCMCA 算法分类结果

第 i 类	分到第 i 类		
	1	2	3
1	2	3	
1	50		
2		45	5
3		5	45

从表中可以看出, 本文提出的 PNPSO-FCMCA 算法较 FCMCA 和 PSO-FCMCA 两种算法的精度

都要高,其精度分别提高了 8.1%和 4.3%,算法效率也分别提高了 17.7%和 7.8%.

表 4 PNPSO-FCMCA 算法分类结果

第 i 类	分到第 i 类		
	1	2	3
1	50		
2		47	3
3		1	49

表 5 三种聚类算法测试时间及精度对比

聚类算法 评价指标	FCMCA	PSO- FCMCA	PNPSO-FCMCA
测试时间/s	5.19	4.63	4.27
聚类精度/%	90.00	93.33	97.33

由于 IRIS 数据集中的第一类数据与其他两类数据是线性分离的,而后两类数据有一定的重叠,因此,采用聚类算法对数据集进行分类时,第一类数据分类精度一般都很高,而第二类和第三类数据的分类结果会因为聚类算法的不同而不一致.这是因为聚类算法所要解决的是如何使样本与聚类中心的紧密程度最高,而聚类算法的区别主要在于迭代寻优的方法不同.由于本文提出的并行小生境粒子群优化算法具有全局搜索能力,能实现样本与聚类中心的最佳匹配,因此分类精度高.而常规的模糊聚类算法因为容易陷入局部最优,导致寻优过程会过早结束而降低分类精度.

4 结束语

本文提出的并行小生境粒子群优化的模糊聚类算法采用山谷函数识别小生境,避免了小生境半径设置不当导致整个小生境算法性能变差,采用惩罚

函数实现共享机制,保证了种群的多样性,避免了算法陷入局部最优,引入聚类有效性评价函数,通过自适应地调整聚类数的取值,从而获得最佳聚类数,避免在聚类数目选取上存在的主观性.仿真结果表明,PNPSO-FCMCA 相比 FCMCA 和 PSO-FCMCA 两种算法的分类精度和算法效率都得到了提高.

参考文献:

- [1] 韩富春,廉建鑫,高文军,等.改进 PSO 与模糊聚类相结合的变压器故障诊断[J]. 电气技术, 2011(5): 1-4.
- [2] 李玲玲,辛浩. FCM 算法及其有效性度量方法[J]. 安徽电子信息职业技术学院学报, 2011, 10(5): 8-10.
- [3] Ursem R K. Multinational evolutionary algorithms[C]// Proceedings of Congress of Evolutionary Computation. Washington, 1999: 1640-1663.
- [4] 刘云龙,林宝军. 一种基于小生境技术的群智能粒子滤波算法[J]. 控制与决策, 2010, 25(2): 316-320.
- [5] 邱学芹. 模糊聚类算法及其聚类有效性的研究[D]. 青岛: 青岛理工大学, 2010.
- [6] Bensaid A M. Validity-guided clustering with applications to image segmentation[J]. IEEE Transactions of Fuzzy Systems, 1996, 4(2): 76-85.
- [7] Kwon S H. Cluster validity index for fuzzy clustering [J]. Electronics Letters, 1998, 34(22): 2176-2177.
- [8] Xie X L, Beni G A. Validity measure for fuzzy clustering[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 1991, 13(8): 841-846.

作者简介:

常天庆 男,(1963—),教授,博士生导师.研究方向为装备自动化系统检测与故障诊断.

李 勇 男,(1983—),博士研究生.研究方向为装备智能故障诊断与健康管理.