一种改进的小生境遗传聚类算法①

孙红艳 王英博 (辽宁工程技术大学 电子与信息工程学院 辽宁 葫芦岛 125105)

摘 要: 传统的遗传算法具有早熟收敛和后期收敛速度慢的缺点，采用改进的小生境技术解决这一问题，同时根据具体问题改进了遗传算子，并将改进后的小生境遗传算法应用于聚类挖掘中。由于聚类挖掘算法中的 K-means 算法对初始值 K 的选取敏感，选取值的不同会导致聚类结果的不同，很容易陷入局部最优，使得聚类结果很差。因此，将改进的小生境遗传算法和 K-means 算法相结合，得出一种改进的小生境遗传聚类算法。验证表明优该算法对提高聚类分析质量是有效的。

关键字: 小生境技术；聚类挖掘；K-means 算法；小生境遗传算法

An Improved Niche Genetic Clustering Algorithm

SUN Hong-Yan, WANG Ying-Bo

(School of Electronic and Information Engineering, Liaoning Technical University, Huludao 125105, China)

Abstract: The traditional genetic algorithm has the shortcomings of premature convergence and slow convergence. This paper adopts improved niche technology to solve this problem. It also uses the specific issues to improve the genetic operators, and the improved niche genetic algorithm is applied to Clustering Mining. As the K-means algorithm in the clustering algorithm for mining has the problem of the selection of the initial value of K-senstive and if we select a different value, it will lead to a different clustering result. It is easy to fall into local optimum. So it will make poor clustering results. Therefore, this article combines the improved niche genetic algorithm with K-means algorithm to produce a new improved algorithm named an improved niche genetic clustering algorithm. It is verified that the algorithm is valid in improving the quality of clustering analysis.

Keywords: niche technology; clustering mining; K-means algorithm; niche genetic algorithm

1 引言

传统遗传算法是模拟生物在自然环境中的遗传和进化过程而形成的一种自适应全局优化搜索算法。生物的进化过程主要是通过染色体之间的交叉和变异来完成的，与此相对应，遗传算法中最优解的搜索过程也模仿生物的进化过程，使用遗传操作数作用于群体进行遗传操作，从而得到新一代群体，其本质是一种求解问题的高效并行全局搜索算法。但是传统的遗传算法也存在一些缺陷，早熟收敛和后期收敛速度慢是影响其应用的两个主要问题。基因缺失被认为是造成早熟收敛的主要原因,采用多样度、成熟度可以标志种群的早熟状况,利用基因补偿、自适应交叉和变异率等技术可以有效防止早熟[1,2]。防止早熟的另一种有效技术是小生境技术[3-5],这是模拟生态平衡的一种仿生技术,即有限的生存资源只能容纳有限的生物数量.这种算法适用于多峰函数的优化计算.有多种实现方法,其中之一是利用排挤思想,即限制相似个体的数量。共享函数和罚函数方法是这一思想的具体实现[6]。虽然小生境技术有很大的优点，但是其本身也有一定的缺点，小生境初始群体的生成是随机的，这会对结果有一定的影响，本文针对这个弱点进行了改进。

聚类是一种有效的数据挖掘方法，其算法有很多，

① 收稿时间:2009-01-19

比较典型的有 K-means 算法。K-means 算法具有简单和高效性，在聚类中占有重要地位。该算法要根据事先确定的 K 值，把待聚类样本分为 K 类，使聚类域中所有样本到聚类中心的距离平方和最小。由于以上优点，K-means 聚类算法已经应用到各种领域，包括图像和语音数据压缩、模式识别、数据分析等。本文将改进的小生境遗传算法来优化 K-means 算法，以达到更好的聚类效果。

2 小生境技术

2.1 小生境技术的基本思想

在自然界中，“人以群分,物以类聚” 是一种正常的自然现象,生物体总是趋向于与自己特征、性状相似的生物生活在一起，进行交配、繁殖。这种交配方式在生物遗传进化过程中，起到了积极的作用。在生物学上，小生境(niching)是指在特定环境中一种组织的功能，而把有共同特性的组织称为物种。小生境技术就是将每一代遗传个体分成若干类，每个类中选出若干适应度较大的个体作为一个类的优秀代表组成一个种群，再在种群中以及不同的种群间通过杂交、变异产生新一代个体群。基于这种小生境的遗传算法由于可以更好的保持解的多样性，已在复杂的多峰值函数求解过程中得到了一定的应用。但是小生境初始群体的产生是随机的会影响算法的性能，因此本文提出一种新的简单策略。

2.2 小生境技术的更新策略

针对初始群体的产生是随机的这一缺点，本文采取新的策略来改善这一缺点。小生境群体中的个体都具有相似性，因此我们将群体中的所有个体 x_i (0<i<m),根据其适应度函数 f(x)计算出每个个体 x_i 相对应的适应度，按照适应度大小降序排列，然后按照给定的小生境容量 w，从前往后按顺序划分出小生境群体。这样小生境内的个体的适应度大小都差不多，很相似。

3 K-means算法的基本思想

K-means 算法以 k 为输入参数，把 n 个对象的集合分为 k 个簇，使得簇内的相似度高，而簇间的相似度低。簇的相似度是关于簇中对象的均值度量，可以看作簇的质心(centroid)或重心(center of gravity)。算法首先随机选择 k 个对像，每个对像初始地代表了一个簇的平均值或中心，对剩余的每个对象根据其与各个簇中心的距离，将它赋给最近的簇，然后重新计算每个簇的平均值，不断重复该过程，直到准则精心策划收敛。准则函数如下：

$$E=\sum_{i=1}^{k}\sum_{x\in C_i}\left|x-\overline{x_i}\right|^2 \tag{1}$$

K-means 算法的描述如下：

① 任意选择 k 个记录作为初始的聚类中心。

② 计算每个记录与 k 个聚类中心的距离，并将距离最近的聚类作为该点所属的类。

③ 计算每个聚集的质心(聚集点的均值)以及每个对象与这些中心对象的距离，并根据最小距离重新对相应的对象进行划分。重复该步骤，直到式(1)不再明显地发生变化。

4 改进的小生境遗传算法与K-means聚类算法相结合

本文将改进的小生境遗传算法应用到聚类挖掘中，把小生境遗传算法的全局优化能力与聚类分析的局部优化能力相结合来优化聚类分析算法的局部性。K-means 算法对初值 K 的选取敏感，会影响到聚类结果的质量，因此，在种群进化中，引入 K-means 操作，同时为了避免早熟现象，采用小生境技术，让种群中的个体不是聚集在一种环境中，而在不同特定的生存环境中进化。这样可以使算法在整个解空间中搜索，以找到更多的最优个体，避免在进化后期适应度高的个体大量繁殖，充斥整个解空间，导致算法停止在局部最优解上。该算法具体步骤如下。

4.1 染色体编码的构成

由于聚类算法具有多维性、数量多等特点，聚类问题的样本数目一般远大于其聚类数目，因此采用基于聚类中心的浮点数编码，将各个类别的中心编码为染色体。这种基于聚类中心的编码方式缩短了染色体的长度，提高了遗传算法的速度，对求解大量数据的复杂聚类问题效果较好。举例：若某一个优化问题含有 5 个变量 x_i (i=1,2,…5),每个变量都有其对应的上下限 $\left[U^i_{\min},U^i_{\max}\right]$,则 X:

5.50	6.80	3.70	3.60	5.00

就表示一个体的基因型其对应的表现型是：$x=[5.50,6.80,3.70,3.60,5.00]^T$。

4.2 初始群体的获得

为了获得全局最优解，初始群体完全随机生成。先将每个样本指派为某一类作为最初的聚类划分，并计算各类的聚类中心作为初始个体的染色体编码串，共生成 m 个初始个体，由此产生第一代种群。

4.3 适度度函数的选取

适应度通常用来度量群体中各个体在优化计处中可能达到或接近于最优解的优良程度。本算法采用式(1)构适应度函数，同于式(1)的值越小说明聚类结果越好，越大说明聚类结果越差，因此选择如下的适应度函数[7]：

$$F=b/(1+E) \qquad (2)$$

其中，b 为常数，可以根据具体问题作调整。E 为聚类准则函数，即上面式子(1)。

根据各个个体的适应度大小对其进行降序排列记忆前 n 个个体(n<m)。

4.4 遗传算子的构成及改进

传统的遗传算法存在着早熟收敛和后期收敛速度慢的弱点，小生境遗传算法是一种新颖的遗传算法。在遗传算法的操作算子中，选择算子起到启发进化方向的作用，交叉算子起到全局搜索的作用，而变异算子通常被认为是一种背景操作或辅助操作，它能够以大于 0 的概率找回丢失的优良基因。

4.4.1 选择算子的构成

本文采用两种选择算子。在小生境中，我们采用$(\mu+\lambda)$选择机制，它被认为是进化算法几种流行的选择机制中选择压最高的一种。当在种群中进行随机配对的交叉操作时，$(\mu+\lambda)$选择机制能产生最快的局部收敛速度。$(\mu+\lambda)$选择策略是指在 μ 个父代个体和由这 μ 个个体交叉产生的 λ 个子个体中选择 μ 个最佳个体。在整个的大群体中本文采用最优保存策略。

4.4.2 交叉算子的构成

因为个体的染色体编码是浮点数编码方法 ，所以交叉操作采用算术交叉算子，即由两个个体的线性组合而产生出两个新的个体。例如：假设在两个个体 X_A^t、X_B^t 之间进行算术交叉，则交叉后所产生出的两个新个体是：

$$X_A^{t+1}=\alpha X_B^t+(1-\alpha)X_A^t$$
$$X_B^{t+1}=\alpha X_A^t+(1-\alpha)X_B^t$$

式中，α 为一参数，这可以是一个常数，也可以是一个由进化代数所决定的变量，分别称为均匀算术交叉和非均匀算术交叉。

4.4.3 变异算子的改进

本文采用 2 种变异算子[8]。第一种是均匀变异，该变异算子运用在算法的初期运行阶段，它使得搜索点可以在整个搜过空间内自由地移动，从而可以增加群体的多样性，使算法处理更多 多的模式。例如：假设有一个体为 $X=x_1x_2...x_k...x_l$，若 x_k 为变异点，其取值范围为 $[U_{\min}^k,U_{\max}^k]$，在该点对个体 X 进行均匀变异操作后，可得到一个新的个体 $X=x_1...x_2...x'_k...x_l$，其中变异点的新基因值是 $x'_k=U_{\min}^k+r\cdot(U_{\max}^k-U_{\min}^k)$：式中，r 为[0，1]范围内符合均匀概率分布的一个随机数。第二种采用自适应调整算随机变异，该算子用在算法的后期运行阶段。但当变异的个体为子种群中的最佳个体时，对该最佳个体及由其变异所得新个体进行(1+1)选择以保证最优个体以概率 1 保留到下一代。

4.5 引入 K-means 操作

先以变异后产生的新群体的编码值为中心，把每个数据点分配到最近的类，形成新的聚类划分。然后按照新的聚类划分，计算新的聚类中心，取代原来的编码值[7]。由于 K-means 算法具有较强的局部搜索能力，因此引入 K-means 操作后，遗传算法的收敛速度可以大大提高。

4.6 进化结束条件

进化代数初始化为 0，每进化一次，即循环一次，代数加 1，若当前进化代数小于规定的代数，则继续进化，否则结束进化。

4.7 本文算法步骤描述

用改进的小生境遗传算法优化后的 K-means 算法步骤描述如下：

STEP1：设置进化代数计数器，随机生成规模为 m 的初始种群；

STEP2：计算个体适应度值并降序排序； //根据排序选择相邻若干个个体进入一个小生境

STEP3：While(不满足结束条件)

STEP4：计算大种群个体方差 D，若 D≥σ 则设置子种群规模 n(n<m，是关于 D 的一个函数)，否则 n=2；依据本文中的小生境新策略找出 n 个个体形成子种群；//自适应策略的关键步骤，子种群规模的确定随种群多样性的变化而自适应变化。根据种群多样性的变化通过阈值 σ 控制实时确定子种群规模。

STEP5：子种群内个体进行算术交叉，并进行(μ

+λ)选择;

STEP6: 用本文采用的两种变异算子进行变异，若变异个体为最优个体则进行(1+1)选择;

STEP7: End While

STEP8: 计算新一代群体的适应度，以最大适应度的最佳个体为中心进行 K-means 聚类。

STEP9: 输出结果。

5 实验结果与分析

本文对原始 K-means 算法，传统遗传算法优化的 K-means 算法以及本文的算法进行了聚类挖掘对比验证。实验数据为一组合成数据和两组实际数据。合成数据集为二维分布数据集(0,0)，(0,1)，(1,0)，(1,1)，(2,1)，(1,2)，(2,2)，(3,2)，(1,4)，(1,5)，(1,6)，(2,4)，(2,5)，(6,6)，(6,7)，(7,6)，(7,7)，(7,8)，(8,6)，(8,7)，(8,8)，(8,9)，(9,7)，(9,8)，(9,9)用 Matlab 仿真共分为三类。数据分布如图 1 所示。

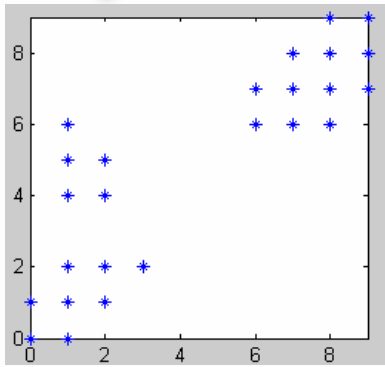

图 1 二维数据聚类结果

实际数据来自 UCI machine learning repository 数据库,所选的数据集是著名的 Iris 数据集和 glass 数据集，在 Turbo C 环境下导入数据进行实验。根据表 1 的试验结果分析，K-means 算法的初始聚类中心的选取对聚类结果影响较大，得到的聚类最优解最差，准确率也最低，而传统遗传算法优化的 K-means 算法相对较好一些，但是准确率也不是很高。相比之下，本文算法得到的聚类结果是最好，准确率也是最高的。

6 结语

本文针对传统遗传算法早熟收敛和收敛速度慢的缺点采用了改进的小生境技术，并且概据具体问题改进了遗传算子，通过阈值实时控制子种群规模，并将改进的小生境遗传算法应用于聚类分析中，针对 K-means 聚类算法对初始值 K 选取的敏感问题，把小生境遗传算法和 K-means 聚类算法相结合。该算法采用基于聚类中心的编码方案，减少了染色体的编码长度，使得聚类效果更好。

表 1 算法性能测试结果

聚类算法	K-means		传统遗传算法优化的K-means		本文算法	
数据集	Iris	glass	Iris	glass	Iris	glass
E的平均值	78.94125	394.05591	78.94056	394.05418	78.93259	394.00156
E的最优解	78.94038	337.02549	78.94012	336.04890	78.930456	335.789452
聚类个数	3	6	3	6	3	6
算法的准确率	79.5%	80.1%	82.2%	83.1%	89.9%	90.5%

参考文献

1 Tian FH, Zhou CG, Tian LH. Prevention againstallel lack in genetic algorithm. Mini-Micro Systems, 2000, 21(9) 947－949.

2 Li SQ, Zhao LY, Shi ZX, et al. An effective method of preventing prematurity of genetic algorithm. Sytems Engineering--Theory &Practice, 1999,19(5):72－76.

3 Zhou M, Sun SD. Principle of genetic algorithm and itsapplications. Beijing: Defense Industry Press, 1999,6.

4 Hao X, Li RH. Multi-population genetic algorithm forcomplex function optimization. Control and Decision, 1998,13(3):263－266.

5 Yu SY, Guo GQ. A class of niche used in genetic algorithms for improving efficiency of searching global optimum. Information and Control, 2001,30(6):526－530.

6 李金屏,李素,杨波.基于小生境算法和聚类分析的快速收敛遗传算法.小型微型计算机系统,2004,25(6):975－978.

7 赖玉霞,刘建平,杨国兴.基于遗传算法的 K 均值聚类分析.计算机工程, 2008,34(20):199－202.

8 周明,孙树栋.遗传算法原理及应用, 2002.