文章编号:1671-4598(2018)04-0136-03 DOI:10.16526/j.cnki.11—4762/tp.2018.04.036 中图分类号:TP273 文献标识码:A

基于人工蜂群优化的K均值聚类算法

廖伍代，朱范炳，王海泉，孙雪凯
(中原工学院 电子信息学院，郑州 450007)

摘要：为了改善K均值聚类算法对初始聚类中心敏感和易于陷入局部最优的不足，提出人工蜂群算法和K均值聚类算法相结合的想法，即基于人工蜂群优化的K均值聚类算法；通过全局寻优能力强的人工蜂群算法初始化K均值的聚类中心并优化聚类中心的位置，从而帮助K均值跳出局部极值，优化聚类效果；将混合聚类算法用Iris、Red Wine和New Red Wine数据集做聚类测试，结果表明该算法既克服了原始K均值聚类算法容易受初始聚类中心影响和不稳定的缺点，又具有良好的性能和聚类效果。

关键词：聚类分析；K均值算法；人工蜂群算法；聚类中心；优化

A K—Means Clustering Algorithm Based on Artificial Bee Colony Optimization

Liao Wudai, Zhu Fanbing, Wang Haiquan, Sun Xuekai
(School of Electric and Information Engineering, Zhongyuan University of Technology, Zhengzhou 450007, China)

Abstract: In order to improve the shortcomings of K—Means algorithm, which are sensitive to initial clustering centers and easily caught in local optimum, proposes an idea that combines K—Means clustering algorithm with artificial bee colony algorithm. That is a K—Means clustering algorithm based on artificial bee colony optimization. With the strong ability of global optimization, the artificial bee colony algorithm can initialize the K—Means clustering centers in an effective way, and move the clustering centers to better positions. As a result of helping K—Means escape from local optimum and optimize clustering effect. Testing the hybrid clustering algorithm with UCI Iris, Red wine and New Red Wine data sets, results show that the algorithm not only overcomes instability of original K—Means, but also provides a better clustering performance.

Keywords: clustering analysis; K—Means clustering; artificial bee colony algorithm; clustering centers; optimization

0 引言

随着科学技术的巨大进步，社会经济也取得了迅速的发展，与人们生活和工作相关的各个领域也包含了越来越多的信息。人们在实际生活和工作中会频繁地面临拥有大量繁杂的数据信息而无法有效、准确提取有价值信息的尴尬境地，数据挖掘理论与技术则应运而生。数据挖掘是从大量数据中挖掘有趣模式和知识的过程，获取有用信息、隐含联系和潜在规律，以引导人们发现大量数据中的有趣信息和规律，以及帮助科研人员进行决策分析。聚类分析是数据挖掘的一种重要技术手段，它是一个把数据集对象或观测划分成若干子集的过程，是一种无监督的学习方法[1]。聚类分析要求聚类划分的类内对象相似度大，而类间对象的差异性大。K均值聚类是经典的聚类算法之一，常用欧氏距离作为相似度的标准进行聚类划分，类内距离和越小表示生成的聚类结果越紧凑和类间越独立，聚类效果越好[2]。K均值聚类算法划分简单易行，收敛速度快；但是，K均值聚类不一定收敛于全局最优解，常常收敛于局部最优，聚类结果常常出现不稳定的现象，影响聚类效果，从而影响人们获取数据中的有效信息和决策的制定。许多研究者采用进化算法和群体智能等方法与K均值混合聚类，包括文献[3-5]分别从初始化聚类中心和优化聚类中心位置的角度考虑，并改进K均值聚类。考虑到人工蜂群算法的诸多优点，与K均值结合，可以同时从初始聚类中心和更新聚类中心位置改进标准K均值算法。

人工蜂群算法是一种模拟自然界蜜蜂采蜜寻找优良蜜源行为的元启发式算法，优化结果不受初始值影响。与其他智能算法相比，人工蜂群算法的优点在于参数设置少，逻辑性好，计算简单易于实现；以较大概率跳出局部极值，具有全局收敛性，鲁棒性强；因为人工蜂群算法是一种通用性强的并行性优化算法，可同时寻优多个解，因此是组合优化和数值优化问题的有效优化工具，具有良好的理论应用和工程应用基础及价值，吸引了众多学者的关注与研究工作。

本文综合考虑人工蜂群算法与K均值聚类算法的优缺点，提出一种基于人工蜂群优化的K均值聚类算法。论述了该混合聚类算法的原理以及实现过程，设计了算法程序流程图；最后，用Iris、Red Wine、New Red Wine数据集做聚类测试与验证，分析实验结果并得出结论。

1 相关算法简介

1.1 K均值（K—Means）聚类算法

在聚类分析问题中，给定采样数据集 $X=\{x_1,x_2,\ldots,x_n\}$，其中 $x_i(i=1,2,\ldots,n)$ 是 d 维数据，即 $x_i\in\mathbf{R}^d$，表示数据集 X 中的每个数据有 d 个属性，$x_i=(x_{i1},x_{i2},\ldots,x_{id})$。K均值聚类算法要将数据集 X 划分成 k 个簇类，即 $C=\{C_1,C_2,\ldots,C_k\}$ 类，每一个聚类的中心为 $c_j(j=1,2,\ldots,k)$。k 个聚类须满足这些条件：$C_j\neq\phi$，$\bigcup_{j=1}^{k}C_j=X$，$i\neq j$，$C_i\cap C_j=\phi$。按照这样的划分条件，在聚类结果中，同一聚类的对象尽可能相似，而不同聚类中的对象尽可能有较大差异，以便于识别与区分。

这里，事先没有给出每一个聚类的标签，因此这是一个无监督学习的问题。数据集 X 中的对象 x_i 与聚类中心 c_j 的欧氏

收稿日期：2017-05-31； 修回日期：2017-06-27。

作者简介：廖伍代(1963-)，男，湖北武汉人，教授，硕士研究生导师，主要从事非线性系统控制，神经网络，智能控制等方向的研究。

距离用 $\| x_i - c_j \|$ 表示。算法的运行步骤描述如下：

1）随机初始化 k 个聚类中心点：$c_1, c_2, \ldots, c_k \in \mathbf{R}^d$；

2）重复循环执行以下操作，直至算法收敛：

对于数据集中的每一个对象 x_i：

$$C_j := \arg\min_j \| x_i - c_j \|^2 \tag{1}$$

对于每一个聚类中心 c_j：

$$c_j = \frac{\sum_{i=1}^{m} 1\{C_i = j\} x_i}{\sum_{i=1}^{m} 1\{C_i = j\}} \tag{2}$$

在以上的算法模型中，步骤 1 表示在数据集 X 中随机选择 k 个对象作为初始聚类中心。公式（1）把每一个数据对象 x_i 划分到离它最近的聚类中心 c_j 对应的类别中；公式（2）把聚类中心 c_j 移动到当前聚类中所有点的均值处。

K 均值聚类算法在一定程度上是收敛的，对聚类公式作如下变形：

$$J(C_j, c_j) = \sum_{i=1}^{m} \| x_i - c_j \|^2 \tag{3}$$

$J(C_j, c_j)$ 描述的是当前聚类划分中的数据对象到对应聚类中心的欧氏距离平方和。K 均值聚类效果用衡量标准函数 E 评价[6]，衡量标准函数为：

$$E = \sum_{i=1}^{k} \sum_{x_i \in C_j} dist(x_i - c_j) \tag{4}$$

式中，E 是所有类内欧氏距离之和，$dist(x_i - c_j)$ 描述的是数据对象 x_i 到其所属聚类中心 c_j 之间的欧氏距离。E 越小表示聚类的划分结果越紧凑和独立，聚类的效果也越好。同时，衡量标准函数 E 可作为人工蜂群算法的优化目标函数，为两种算法的结合提供切入点。

1.2　标准人工蜂群（ABC）算法

人工蜂群算法的基本要素包括蜂群、蜜源和蜜源适应度[7]，算法将蜂群分为采蜜蜂、观察蜂和侦察蜂 3 种。ABC 算法一般通过较大的适应度值引导算法向全局最优进化[8]，对于最大值优化问题，可用待优化问题的目标函数 f 表示适应度函数 fit；对于最小值优化问题，适应度函数用式（5）表示：

$$fit = \begin{cases} \dfrac{1}{1+f}, & f \geqslant 0 \\ 1 + abs(f), & f < 0 \end{cases} \tag{5}$$

ABC 算法步骤叙述如下：

蜂群的初始化产生解：ABC 算法的初始化阶段，包括种群规模 SN，最大迭代次数 MCN，开采度次数 $Limit$。蜂群中蜜蜂数量和食物源数量相等，且所有蜜蜂都是侦察蜂模式。通过式（6）随机产生 SN 个解并计算其适应度，将适应度按由大到小的顺序排列，前一半作为采蜜蜂，后一半作为观察蜂和侦察蜂。

$$x_{id} = x_{id\min} + rand(0,1)(x_{id\max} - x_{id\min}) \tag{6}$$

对于任一解 x_i 的任一分量 $x_{id}(d = 1,2,\ldots,D)$ 都进行初始化，$x_{id\min}$ 代表可行解空间分量的最小值，$x_{id\max}$ 代表可行解空间分量的最大值。

采蜜蜂搜索阶段：采蜜蜂在初始阶段的蜜源附近，通过方程（7）搜索产生一个新解，即为候选蜜源进行开采。

$$v_{id} = x_{id} + rand(-1,1)(x_{id} - x_{jd}) \tag{7}$$

式中，$j \in \{1,2,\ldots,N\}$，$j \neq i$ 表示在 N 个蜜源中随机选取一个不同于 x_i 的蜜源。计算新解的适应度 fit_i 并进行适应度大小评价，在 v_i 和 x_i 之中采用贪婪策略进行选择[9]。

观察蜂跟随阶段：所有采蜜蜂完成搜索之后，采蜜蜂会把蜜源信息及适应度分享给观察蜂。观察蜂通过选择概率 P_i 决定每只采蜜蜂被跟随的概率

$$P_i = \frac{fit_i}{\sum_{k=1}^{N} fit_k} \quad or \quad P_i = \frac{fit_i}{\max(fit_k)} \tag{8}$$

观察蜂根据选择概率，采用轮盘赌策略选择采蜜蜂跟随；轮盘赌成功，跟随采蜜蜂并再次更新其对应的蜜源。若新蜜源对应解的适应度比之前的好，观察蜂会将新解保存；反之，观察蜂将会保留原来的解，同时解的迭代搜索次数 iteration 会加 1。

侦察蜂阶段：如果某一食物源在被搜索开采 Limit 次之后仍没有被更新，相应的采蜜蜂和观察蜂则会放弃该蜜源，转换为侦察蜂模式，按照公式（6）进行全局随机搜索，寻找一个新的蜜源代替被舍弃的蜜源。然后返回到采蜜蜂的搜索阶段，3 种蜜蜂依次进行工作，重复循环搜索，最终找到待优化问题的最优解。

1.3　人工蜂群算法的改进

对人工蜂群算法的改进研究，主要从提高收敛精度和加快收敛速度两个方面进行。在标准人工蜂群算法中，主要采用贪婪选择和轮盘赌策略来选择新解和选择采蜜蜂进行跟随，但是两种方法都是依靠个体蜜源的适应度进行选择。这种做法的结果是过度贪婪选择适应度高的蜜源，解的多样性丢失，放弃了具有开发价值的解，最终导致收敛精度降低。本文在应用人工蜂群算法时，为了提高收敛精度，做了算法的改进工作。首先，采用比例选择的方法代替轮盘赌策略，比例选择可以使每一个采蜜蜂都有机会得到跟随，进而使每一个解都会得到进一步开发；当然适应度高的蜜源，根据比例选择会被更多的观察蜂跟随，对应解的开发力度会更大；但是测试结果并不好，最终的选择概率会平均分布，目标函数值会稳定在精度更低的优化结果处。因此，轮盘赌在蜂群算法中还是一种优秀的策略。轮盘赌策略是采用一个（0，1）之间的随机数与选择概率 P_i 进行比较，若随机数小于 P_i，观察蜂跟随采蜜蜂；否则，观察蜂不跟随，继续比较下一个采蜜蜂的选择概率，直到所有观察蜂都进行了跟随工作。为了使选择概率不完全依靠个体蜜源的适应度值，设计新的选择概率计算公式如下：

$$P_i = \frac{0.9 fit_i}{\max(fit_k)} + 0.1 \tag{9}$$

这是一种较为有效的改进方法，测试结果优于标准人工蜂群算法。

2　基于人工蜂群优化的 K 均值聚类实现方法

2.1　食物源的编码与更新方法

根据聚类分析的问题背景，建立样本数据集 $X = \{x_1, x_2, \ldots, x_n\}$，其中 $x_i(i = 1,2,\ldots,n)$ 是 d 维数据，说明每个数据对象有 d 个属性。K 均值算法将数据集 X 划分成 k 个簇，可设置人工蜂群算法要素，初始化蜂群有 $SN/2$ 个采蜜蜂，每个蜜蜂代表一个数据集划分，$Z = \{Z_1, Z_2, \ldots, Z_k\}$，其中 $z_i(i = 1,2,\ldots,k)$ 为 d 维向量是，表示 Z_i 的聚类中心，k 是聚类数目，即每个蜜蜂为 k 个 d 维向量的聚类中心。

蜂群初始化时，从样本数据集中随机选取 k 个数据，每个

数据包含 d 维。因此，每只采蜜蜂对应的食物源是一个 $k \times d$ 的矩阵，将蜂群选取的初始聚类中心送给 K 均值执行聚类步骤，得出聚类中心和聚类衡量函数值 E；接着采用标准蜂群算法中邻域更新方法更新得到的聚类中心位置，继续执行 K 均值聚类算法，并以适应度函数引导收敛；重复操作以上步骤，循环迭代，直到达到算法的终止条件。

2.2 适应度函数设计

在人工蜂群算法中，适应度函数的选取是影响算法稳定性和收敛性的关键因素[10]。在聚类分析问题中，要求类内成员具有较好的相似性，类间成员有较大的差异性；用欧氏距离表示这种相似度，以聚类分析的衡量标准函数描述聚类分析的效果。因此，本文把衡量标准函数作为基于人工蜂群优化的 K 均值聚类算法的目标函数，从而设计适应度函数。每个蜜蜂可以通过式（4）计算出聚类衡量函数 E，所求优化目标是 E 的最小值。因此，适应度函数设计[11]为：

$$fit = \frac{1}{1+E} \tag{10}$$

基于人工蜂群优化的 K 均值聚类算法的程序流程设计如图 1 所示。

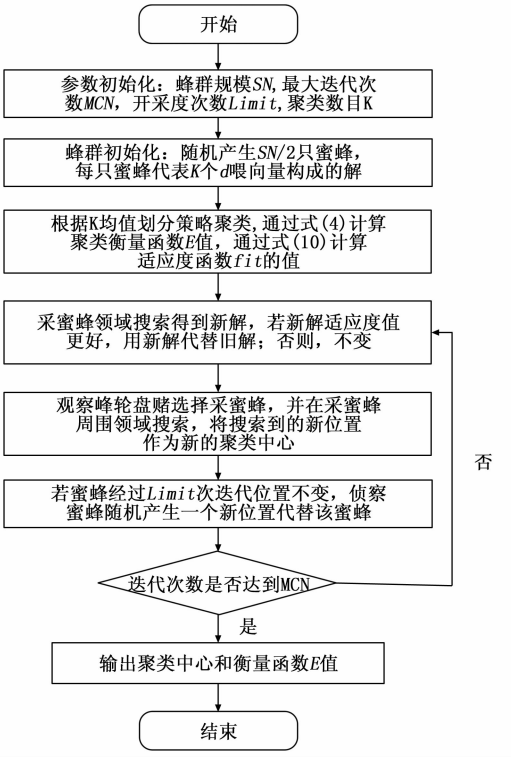

图 1 蜂群优化 K 均值算法流程图

3 实验结果及分析

为了验证本文提出的算法确实改善了聚类效果与精度，利用 Matlab 进行算法仿真。本文采用 UCI 机器学习数据库聚类索引中的 Iris、Red Wine 和 New Red Wine 数据集[12]进行混合聚类算法的测试。实验环境：计算机主机 i3CPU、主频 3.30 GHz，运行内存 2.0G；软件版本为 MATLABR2014a，分别运行原始 K 均值聚类算法和基于人工蜂群优化的 K 均值聚类算法 20 次。Iris、Red Wine 和 New Red Wine 数据集中数据表示的是不同科鸢尾属植物和不同级别红酒各种物质的含量值，根据这些数值划分识别鸢尾属植物的科目和红酒的级别。所采用的数据集详细信息见表 3.1。

在实验中，对于不同的聚类数据集对象，蜂群规模 SN 设置不同，聚类数目设置见表 1；设置蜂群算法的最大循环搜索次数 MCN 为 500，同一蜜源的可重复开采次数为 Limit 为 10，实验结果的相关数据记录如表 2～4 所示。

表 1 实验所用数据集信息

数据集名称	包含样本数	样本属性数	聚类数目
Iris	150	4	3
Red Wine	178	13	3
New Red Wine	1599	11	6

表 2 Iris 数据集聚类结果对比

算法	收敛时间	E 最小值	E 最大值	E 平均值	E 标准差
K 均值	0.0064s	78.9408	145.2793	91.8444	26.4819
ABC－K 均值	2.73s	69.7751	71.2738	70.0771	0.1051

表 3 Red Wine 数据集聚类结果对比

算法	收敛时间	E 最小值	E 最大值	E 平均值	E 标准差
K 均值	0.0078s	2.3707e+6	2.6407e+6	2.4750e+6	0.1312e+6
ABC－K 均值	4.652s	2.2524e+6	2.3029e+6	1.6471e+6	0.0088e+6

表 4 New Red Wine 数据集聚类结果对比

算法	收敛时间	E 最小值	E 最大值	E 平均值	E 标准差
K 均值	0.0305s	1.9341e+5	1.9589e+5	1.9355e+5	570
ABC－K 均值	12.506s	1.8930e+5	1.9095e+5	1.8977e+5	38.2076

分析上表数据可知，原始 K 均值算法收敛速度快，只需很短的时间就能输出聚类结果；但是聚类衡量函数值 E 变化幅度大，说明不同次的实验选择不同的初始聚类中心，对结果影响较大，即原始 K 均值算法容易受初始聚类中心影响，且易于陷入局部最优；原始 K 均值算法相对不稳定，体现在衡量函数的标准差较大。

引入人工蜂群算法优化 K 均值聚类算法，增加了算法的时间复杂度，而且聚类数目越多，聚类对象的属性越多，算法收敛时间会越长。但是，该算法确实克服了原始 K 均值算法的缺点，降低了聚类衡量函数 E 值的变化幅度，改善了聚类效果，使聚类结果更加稳定。在时间复杂度允许的范围内，提高聚类的精度是十分有意义的；可以提高样本数据集在无监督情况下的划分准确率，更好地帮助研究者发现数据间的联系与规律，做出分析与决策。

4 结论

本文提出一种基于人工蜂群优化的 K 均值聚类分析算法。同时从随机初始化聚类中心和优化聚类中心位置入手，利用聚类衡量标准函数作为蜂群优化的目标函数，并以此设计适应度函数，以较大概率引导聚类向全局最优解收敛，找到更优的聚类中心。实验证明，该算法有效克服了 K 均值聚类算法易陷入局部最优和不稳定的缺点，提高了对聚类中离群点和边缘点

（下转第 156 页）