

基于布谷鸟搜索改进的聚类算法

孙伟鹏¹, 孟 斌², 吴锡生¹

(1 江南大学 物联网工程学院, 江苏 无锡 214122;

2 中船重工集团第七〇二研究所 软件工程中心, 江苏 无锡 214082)

摘 要: 在分析 K-均值算法和布谷鸟搜索(CS)算法的基础上, 本文提出了一种基于改进布谷鸟搜索算法的聚类算法. 该算法将局部搜索能力强的 K-均值算法和全局搜索能力强的布谷鸟搜索算法相结合, 既提高了布谷鸟搜索算法的局部搜索能力, 加快了收敛速度. 同时因为布谷鸟搜索算法具有较强的全局搜索能力, 有效地防止了早熟收敛现象的发生, 因此可以有效地获得全局最优解. 实验表明该聚类算法有更好的收敛效果.

关键词: 聚类; k-means 算法; 布谷鸟搜索算法; 收敛速度; 全局最优

中图分类号: TP301.6

文献标识码: A

文章编号: 1000-7180(2018)08-0016-05

Improved Clustering Algorithm Based on Cuckoo Search

SUN Wei-peng¹, MENG bin², WU Xi-sheng¹

(1 School of IoT Engineering, Jiangnan University, Wuxi 214122, China;

2 Software Engineering Center, China Shipbuilding Industry Group No. 702 Institute,
Wuxi 214082, China)

Abstract: Based on the analysis of K-means algorithm and cuckoo search (CS) algorithm, this paper proposes a clustering algorithm based on improved cuckoo search algorithm. The algorithm combines the strong local search ability provided by K-means algorithm with the strong global search ability provided by improved cuckoo search algorithm. Which not only improves the local search ability of the cuckoo search algorithm, but also accelerates the convergence rate. At the same time, because the cuckoo search algorithm has strong global search ability, it can effectively prevent the premature convergence phenomenon, so it can effectively achieve the global optimal result. Experiments show that the clustering algorithm has a better convergence effect.

Key words: clustering; k-means; cuckoo search algorithm; convergence speed; global optimal

1 引言

聚类分析是数据挖掘领域的一个重要研究方向^[1]. K-均值算法作为基于划分的经典聚类方法^[2-3]之一, 但该算法存在对初始聚类中心点的选取比较敏感, 聚类结果不稳定等缺点^[4-8].

布谷鸟搜索(Cuckoo Search CS)算法^[9]是一种新型元启发算法, 它采用 Lévy 飞行产生随机步长, 具有很强的全局搜索能力. 但 CS 算法同其他一些种群智能算法一样, 也存在后期搜索速度慢、精度不

高等缺点^[10-14]. 因此, 陆续有一些学者对原始 CS 算法进行了改进, 如动态自适应步长的 CS 算法、按维度改进 CS 算法等^[15-16]. 但没有结合聚类算法特点进行改进. 因此, 本文在对传统 CS 算法进行改进的基础上, 利用 K-均值算法本身具有很强的局部搜索能力和每次迭代得到新的中心点更加接近局部最优的特点, 提出一种新的基于布谷鸟搜索改进的 K-means 聚类(BICS-K-means)算法. 该算法能够在全局搜索和局部搜索之间保持很好地平衡, 同时提高 CS 算法的搜索精度和收敛速度.

2 相关算法

2.1 K-均值聚类算法

K-均值算法是一种基于划分的经典聚类算法.该算法流程如下:

- (1) 从 n 个数据样本中任意选取 k 个样本作为初始的聚类中心点;
 - (2) 计算每个样本与聚类中心点的距离,并根据最小距离将样本进行划分为 k 个集合;
 - (3) 根据各集合中的样本采取平均值的方法计算新的聚类中心点;
 - (4) 计算样本误差函数:
- $$G = \sum_{j=1}^k \sum_{i=1}^{n_j} (x_i^{(j)} - m_j)^2 \quad (1)$$
- (5) 若样本误差函数值 G 的取值收敛,则返回聚类中心,算法结束.否则转至步骤(2).

2.2 标准布谷鸟搜索算法

布谷鸟在自然界中通过随机或类似随机的方式搜索适合自己产卵的宿主鸟巢,为抽象出布谷鸟搜索宿主鸟巢的行为,需要设定3个假设条件:

- (1) 每只布谷鸟每次只产一个蛋,并随机选择一个鸟巢来孵化它.
- (2) 在随机选择的一组鸟巢中,最好的鸟巢会被保留到下一代.
- (3) 所有布谷鸟每次选取宿主鸟巢总数 n 是固定的,每个宿主鸟巢主人发现一枚外来鸟蛋的概率固定为 $p_a \in [0, 1]$.

在这3个假设条件基础上,布谷鸟搜索鸟巢位置的路径更新公式为

$$x_i^{t+1} = x_i^t + \alpha(x_i^t - x_{\text{best}}^t) \oplus L(\lambda) \quad (2)$$

式中, x_i^t 表示第 t 代第 i 个解; x_{best}^t 表示第 t 代中的最优解; $\oplus$ 表示点对点乘法; α 表示步长控制量,用于控制随机搜索的范围,在标准 CS 算法中是常数; $L(\lambda)$ 表示服从参数 $(1 < \lambda \leq 3)$ 的 Lévy 分布产生的一个随机搜索向量,如式(3)所示.

$$L(\lambda) \sim u = t^{-\lambda} (1 < \lambda \leq 3) \quad (3)$$

每次鸟巢位置更新后,生成随机数 $r \in [0, 1]$ 与发现概率 P_a 比较,当 $r < P_a$ 时,丢弃部分鸟巢,并采用偏好随机游走重新生成相同数量的新鸟巢,否则不变;最后保留最优的鸟巢位置.

3 改进的布谷鸟搜索(ICS)算法

针对标准 CS 算法,本文主要从两个方面对其进行改进:(1)借鉴遗传算法的思想,在算法进化过

程中加入选择、交叉、变异操作,以进一步增加种群多样性;(2)采用自适应发现概率替换原有的固定发现概率,使得算法在陷入局部最优时更容易生成新个体,避免算法“早熟”.

3.1 为布谷鸟算法加入遗传算子操作

为了进一步增加种群的多样性,同时获得较高质量的个体.本文借鉴遗传算法的思想为布谷鸟搜索算法引入了遗传算子操作.该操作可以利用已有的质量较好的鸟巢产生高质量的鸟巢.具体操作如下:在标准 CS 算法执行完鸟巢更新操作后,首先采用基于轮盘赌选择策略根据适应度的优劣按一定比例随机选择出 n (种群规模大小) 个鸟巢(适应度好的鸟巢可能被多次选中).将选出的 n 个鸟巢随机配对成 $n/2$ 对配对鸟巢,然后进行两两交叉、对交叉后的鸟巢进行变异操作,得到 n 个新的鸟巢.计算得到的新鸟巢的适应度,并与之前的鸟巢进行比较,从所有鸟巢中选择出最优的 n 个鸟巢作为最终种群.

3.2 发现概率自适应

标准 CS 算法中发现概率 P_a 是一固定值,但布谷鸟搜索算法在寻优的过程中,随着鸟巢质量的逐步提高,适当地增加种群进化的强度,可以有效避免算法陷入局部最优解.比较好的 P_a 应当是当算法前后迭代所得的最优解接近时,表明鸟巢进化较充分,需要增加解的多样性来避免“早熟”,此时需要较大的发现概率以提高算法的全局优化性能.当前后迭代所得的最优解相差较大时,表明还有很大进化空间,此时需要较小的发现概率以保证解的充分优化.因此,本文采用式(4)来实现 P_a 的动态变化.

$$P_{a,t+1} = (P_{a,\max} - P_{a,\min}) \sin \left[\frac{\pi}{2} \cdot \left(\frac{f_{\text{best}}^t - f_{\text{best}}^{t-1}}{f_{\text{best}}^{t+1} - f_{\text{best}}^t} \right)^m \right] + P_{a,\min} \quad (4)$$

式中, $P_{a,t+1}$ 表示第 m ($m > 1$) 代发现概率; $P_{a,\max}$ 、 $P_{a,\min}$ 分别表示 P_a 的最小和最大控制参数; f_{best}^t 、 f_{best}^{t+1} 、 f_{best}^{t-1} ($f_{\text{best}}^{t-1} < f_{\text{best}}^t < f_{\text{best}}^{t+1}$) 分别表示第 $t-1$ 、 t 和 $t+1$ 代得到的最优解; m ($m > 1$) 是常数,用于扩大前后迭代最优解的比值范围.

4 基于改进布谷鸟搜索算法的 K-means 聚类(BICS-K-means)算法

4.1 编码和适应度函数

假设原始样本数据集可划分成 k 类,每个样本包含 d 维特征,则算法的编码方法是选取 k 组向量坐标作为一个解.这里鸟巢的位置由 k 个聚类中心

表示,样本向量的维数为 d ,所以鸟巢的位置由 $k \times d$ 维矩阵组成,编码方式如式(5)所示.

$$\mathbf{A} = \begin{bmatrix} \text{nest}_{11} & \text{nest}_{12} & \cdots & \text{nest}_{1d} \\ \text{nest}_{21} & \text{nest}_{22} & \cdots & \text{nest}_{2d} \\ \cdots & \cdots & \cdots & \cdots \\ \text{nest}_{k1} & \text{nest}_{k2} & \cdots & \text{nest}_{kd} \end{bmatrix} \quad (5)$$

式中, $\text{nest}_{11}, \text{nest}_{12}, \cdots, \text{nest}_{1d}$ 表示第 1 个类的聚类中心向量; $\text{nest}_{21}, \text{nest}_{22}, \cdots, \text{nest}_{2d}$ 表示第 2 个类的聚类中心向量; $\text{nest}_{k1}, \text{nest}_{k2}, \cdots, \text{nest}_{kd}$ 表示第 k 个类的聚类中心向量. 本文在 K -均值聚类过程中采用最近距离来确定样本所属类别.

本文采用误差平方和函数值作为鸟巢适应度来进行质量评价. 其表达式如式(6)所示:

$$\text{fitness} = \sum_{j=1}^k \sum_{i=1}^{n_j} (x_i^j - m_j)^2 \quad (6)$$

式中, $m_j (j = 1, 2, \cdots, k)$ 代表各个聚类中心; x_i 是样本数据. 在样本集给定时,fitness 的取值取决于 k 个聚类中心. fitness 描述 n 个样本聚类成 k 类时所产生的误差平方和. 显然,fitness 的取值越小,说明聚类的效果越好.

4.2 改进的布谷鸟搜索算法的鸟巢更新

在标准布谷鸟搜索算法中,每次根据 Lévy 飞行来确定下一代鸟巢飞行的位置. Lévy 飞行随机产生的步长时短,步长越长则搜索精度越低,越容易搜索全局最优;步长越短则搜索精度越高,但会严重影响搜索速度. 因此,采用 Lévy 飞行产生步长虽具有随机性,但缺乏自适应性. 而好的布谷鸟位置搜索策略应该是每次位置更新后鸟巢的适应度强于更新前鸟巢的适应度,且随着鸟巢适应度的不断优化,鸟巢位置更新的步长也应该随之减小来确保搜索的精度. 结合 K -均值算法具有以下特点:(1)算法每次迭代所得到的新的中心点适应度一般都好于之前的中心点适应度;(2)随着迭代次数的增加,每次中心点的位置变化与之前相比较也不断减小,具有很强的局部寻优性. 因此可以使用 K -均值算法中心点的更新来模拟鸟巢位置状态的更新. BICS 算法鸟巢位置更新步骤可描述如下:

(1)以当前鸟巢所代表的 k 个中心点执行一次 K -均值划分操作,通过簇内平均值得到以新的 k 个中心点表示的鸟巢并计算其适应度.

(2)对当前鸟巢采取自适应步长调整更新鸟巢并计算其适应度. 其中自适应步长控制量 α_i 来源于文献[17]. 计算公式如式(7)所示:

$$\alpha_i = \text{step}_{\min} + (\text{step}_{\max} - \text{step}_{\min}) \cdot d_i \quad (7)$$

式中, $\text{step}_{\max}$ 表示预定义步长的最大值; $\text{step}_{\min}$ 表示预定义步长的最小值. d_i 表示步长调整因子,其计算公式如下所示:

$$d_i = \left| \frac{F_i - F_{\text{best}}}{F_{\text{best}} - F_{\text{worst}}} \right| \quad (8)$$

式中, F_{best} 为本次迭代所得到的最优鸟巢适应度值; F_{worst} 为本次迭代所得的最差鸟巢适应度值; F_i 为当前鸟巢的适应度值.

(3)比较(1)和(2)中的适应度,取适应度较好的鸟巢作为候选鸟巢. 比较候选鸟巢和当前鸟巢的适应度,如果候选鸟巢的适应度好于当前鸟巢的适应度,则用候选鸟巢替换当前鸟巢,并更新鸟巢适应度.

由上面 BICS 算法的鸟巢更新操作可知:在改进后的布谷鸟搜索算法迭代过程中,算法不断通过 K -均值操作加快局部搜索和提高搜索精度,同时当鸟巢陷入局部最优时可通过自适应步长跳出局部最优. 因而该算法的搜索速度和寻优精度都有较大的提高.

4.3 BICS-K-means 聚类算法的实现步骤

根据上文分析,BICS-K-Means 聚类算法的实现步骤如下:

(1)给定待聚类样本和聚类数 k . 初始化布谷鸟算法的各个参数,如算法的种群规模 n 、发现概率控制量 $P_{a,\max}, P_{a,\min}$ 、算法的迭代次数 maxIter 、最小步长 $\text{step}_{\min}$ 、最大步长 $\text{step}_{\max}$ 和适应度函数等.

(2)在可行解空间中初始化 n 个鸟巢,对样本数据进行聚类划分,根据式(6) 计算每个鸟巢的适应度值并保留最优鸟巢.

(3)根据 3.2 小节提出鸟巢更新策略更新当前鸟巢.

(4)对种群实施 2.1 小节引入的遗传算子操作,筛选得到适应度最优的 n 个鸟巢.

(5)生成随机数 r 与按照 2.2 小节计算得到的 P_a 对比,若 $r < P_a$,则抛弃该鸟巢并重建新鸟巢;否则,不改变鸟巢.

(6)进行聚类划分并计算鸟巢的适应度,选出最优的鸟巢,与目前所得到的最优鸟巢对比. 若适应度更优,则将此鸟巢设置为最优鸟巢.

(7)如果未达到最大的迭代次数或最优值不满足终止条件则返回(4) 继续执行; 否则输出最优的鸟巢代表的聚类中心点.

5 实验结果与分析

为了验证本文提出算法的有效性. 选用 3 个

UCI 标准数据集 Iris、Wine 和 Haberman 作为实验测试数据,数据集基本信息如表 1 所示.

表 1 实验中涉及的数据集			
数据集名称	样本数目	属性维度	类别个数
Iris	150	4	3
Wine	178	13	3
Haberman	306	3	2

试验环境为:软件环境:操作系统 Mac OS,运行软件为 Python 2.7;硬件环境:CPU Intel 2.40 GHz,内存为 4 GB.实验分别采用原始 K-means 聚类算法、基于改进的遗传算法的 K-means 聚类(GA-K-means)算法、基于改进的粒子群算法的 K-means 聚类(PSO-K-means)算法、基于布谷鸟搜索算法的 K-means 聚类(CS-K-means)算法以及本文提出的基于改进的布谷鸟搜索算法的 K-means 聚类(BICS-K-means)算法在表 1 的数据集上分别进行仿真实验.其中鸟巢规模为 20,迭代次数为 20.五种算法分别在各个数据集上进行 20 次独立重复实验,实验的平均适应度函数值和聚类准确率分别如表 2~3 所示.

表 2 算法的平均适应度函数值			
算法	Iris	Wine	Haberman
K-means	78.945	2 370 945	30 914
GA-K-means	79.940	2 370 745	30 600
PSO-K-means	79.972	2 370 745	30 526
CS-K-means	78.945	2 370 745	30 519
本文算法	78.940	2 370 689	30 507

表 3 算法的平均聚类准确率				%	
数据集	Iris	Wine	Haberman		
K-means	81.8	72.1	52.3		
GA-K-means	87.8	76.3	52.8		
PSO-K-means	89.1	71.4	52.9		
CS-K-means	88.6	72.1	52.6		
本文算法	89.3	72.8	53.5		

为了能更加直观地看到不同算法的收敛性能,绘制各算法的收敛曲线如图 1~3 所示.

从表 2~3 的实验结果可以看出,本文算法与其他四种方法相比能够有效地获取适应度较优的解.同时我们可以看到 GA-K-means 算法在数据集 Wine 上所得到的适应度函数值虽然大于 BICS-K-means,但是它的聚类准确度却高于本文算法.这是因为我们选取的适应度函数值与聚类的准确度间没有完全符合单调递减函数关系,而只是大致上符合.

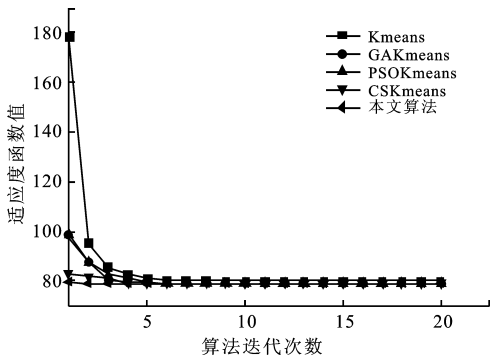


图 1 Iris 数据集收敛曲线

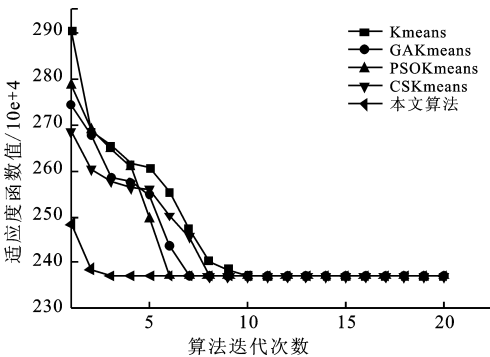


图 2 Wine 数据集收敛曲线

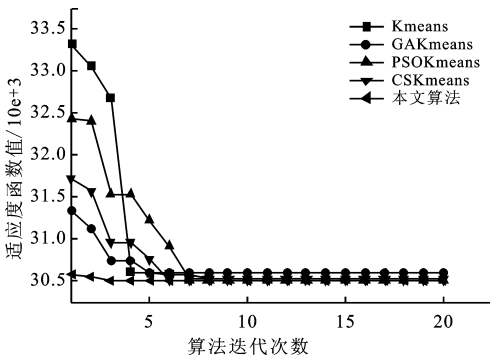


图 3 Haberman 数据集收敛曲线

实验结果可以看出,本文算法在保证获得较高的聚类精度前提下,算法在迭代次数上也有明显的优势.

6 结束语

本文提出了一种改进的布谷鸟搜索算法,并将该算法与 K-means 聚类算法相结合.通过自适应发现概率和对布谷鸟搜索算法引入遗传算子操作,增加了种群解的多样性.在鸟巢位置优化过程中引入 K-means 操作使得算法在全局搜索和局部搜索间相协调,这样不但可以加快搜索速度,而且寻优精度也得到较大的提高.通过试验分析验证了改进后的算法能够获得较高的准确率和稳定性.但该算法也存

在以下不足:(1)为了更加有效地获取全局最优解,需要比较大的种群规模,这时如果使用串型算法,则运行算法需要大量的时间.考虑到各个体间适应度的计算是相互独立的,所以后期可以考虑对该算法进行分布式改进.(2)本文选取的适应度函数并不能很好地与聚类准确率保持一致的单调性.因此,虽然算法可以取得具有最优适应度的鸟巢,但聚类的结果相对于其他一些优秀聚类算法的准确率可能不是最优的,针对这个问题后期还需要找寻更好的适应度函数来改进该算法.

参考文献:

- [1] Han Jiawei, Kamber. 数据挖掘概念与技术[M]. 范明, 孟小峰. 北京: 机械工业出版社, 2008.
- [2] 孙吉贵, 刘杰, 赵连宇. 聚类算法研究[J]. 软件学报, 2008, 19(1): 48-61.
- [3] Xu Jinhua, Liu Hong. Web users clustering analysis based on k-means algorithm[C]//IEEE 2010 international conference on information, networking and automation, Kunming, China, IEEE, 2010.
- [4] 谢娟英, 王艳娥. 最小方差优化初始聚类中心的 K-means 算法[J]. 计算机工程, 2014, 40(8): 205-211.
- [5] Goyal M, Kumar S. Improving the initial centroids of k-means clustering algorithm to generalize its applicability[J]. Journal of the Institution of Engineers, 2014, 95(4): 345-350.
- [6] 邢长征, 谷浩. 基于平均密度优化初始聚类中心的 k-means 算法[J]. 计算机工程与应用, 2014, 50(20): 135-138.
- [7] 陈光平, 王文鹏, 黄俊. 一种改进初始聚类中心选择的 K-means 算法[J]. 计算机系统, 2012, 33(6): 1320-1323.
- [8] Deb S, Yang X S, et al. Towards enhancement of performance of K-means clustering using nature-inspired optimization algorithms[J]. The Scientific World Journal, 2014: 564-829.
- [9] Yang X S, Deb S. Cuckoo search via Lévy flights[C]//NaBIC 2009: Proceedings of the 2009 World Congress on Nature & Biologically Inspired Computing, Piscataway, NJ; 2009: 210 - 214.
- [10] 陆林花, 王波. 一种改进的遗传聚类算法[J]. 计算机工程与应用, 2007, 43(21): 170-172.
- [11] 陶新民, 徐晶. 一种改进的粒子群和 K 均值混合聚类算法[J]. 电子与信息学报, 2010, 32(1): 92-97.
- [12] 刘长平, 叶春明. 具有 Lévy 飞行特征的蝙蝠算[J]. 智能系统学报, 2013, 8(3): 240-246.
- [13] Yang X S, Deb S. Engineering optimization by cuckoo search[J]. Computers & Operations Research, 2013, 40(6): 1616-1624.
- [14] 沈艳, 余冬华, 王浩类. 粒子群 K-means 聚类算法的改进[J]. 计算机工程与应用, 2014, 50(21): 125-128.
- [15] 王李进, 尹义龙, 钟一文. 逐维改进的布谷鸟搜索算法[J]. 软件学报, 2013, 23(11): 2687-2698.
- [16] 张永韩, 汪镭, 吴启迪. 动态适应布谷鸟搜索算法[J]. 控制与决策, 2014, 29(4): 617-622.
- [17] 邓洪清, 周永权. 一种自适应步长布谷鸟搜索算法[J]. 计算机工程与应用, 2013, 49(10): 68-70.

作者简介:

孙伟鹏 男, (1990-), 硕士研究生. 研究方向为聚类分析、数据挖掘. E-mail: sunwp244372610@163.com.

孟 斌 男, (1980-), 高级工程师. 研究方向为数据挖掘.

吴锡生 男, (1959-), 博士, 硕士生导师. 研究方向为聚类分析、数据挖掘、模式识别.

(上接第 15 页)

- [8] 袁威, 郝正航, 张宏俊. 带光储系统的模块化电力电子变压器在配电网中的仿真研究[J]. 电测与仪表, 2016, 53(17): 85-91.
- [9] 唐庚, 徐政, 刘昇. 改进式模块化多电平换流器快速仿真方法[J]. 电力系统自动化, 2014, 38(24): 56-61.
- [10] Zhuo L, Morris G R, Prasanna V K. High-performance reduction circuits using deeply pipelined operators on FPGAs[J]. IEEE Transactions on Parallel & Distributed Systems, 2007, 18(10): 1377-1392.

作者简介:

李雄飞 男, (1993-), 硕士研究生. 研究方向为电磁暂态仿真和硬件加速. E-mail: 448134158@qq.com.

贺光辉 男, (1980-), 博士, 副研究员. 研究方向为通信系统、VLSI、硬件加速等.

项胤兴 男, (1986-), 硕士, 工程师. 研究方向为电力系统仿真.