

混合细菌觅食和粒子群的 k-means 聚类算法

闫 婷, 谢红薇

(太原理工大学 计算机科学与技术学院, 山西 太原 030024)

摘 要: 针对传统 k-means 聚类算法中初值的敏感性, 容易陷入局部最优解的缺陷, 提出了一种优化初始聚类中心的 k-means 聚类算法. 该算法将全局搜索能力强的粒子群算法与局部搜索能力强的细菌觅食算法结合, 将细菌的趋化行为简化为粒子群中粒子寻找最优解的过程, 再利用细菌完成复制、迁徙操作. 将混合算法的最优解确定为初始聚类中心, 解决了 k-means 算法随机选择聚类中心的弊端. 对 Iris、Wine、Glass 等 UCI 数据集的测试结果表明, 该算法的准确率和稳定性都高于流行的聚类算法, 能够更有效地解决复杂的优化问题.

关键词: k-means; 细菌觅食优化算法; 粒子群算法; 聚类算法

中图分类号: TP39

文献标识码: A

文章编号: 1000-7180(2016)06-0059-04

Novel K-means Clustering Algorithm Combining Particle Swarm Optimization and Bacterial Foraging

YAN Ting, XIE Hong-wei

(Department of Computer Science and Technology, Taiyuan University of Technology, Taiyuan 030024, China)

Abstract: This paper, we propose a novel clustering algorithm combining bacterial foraging algorithm (BFO) which has high global search ability with particle swarm algorithm (PSO) which has high local search ability based on the initial values of traditional k-means clustering algorithm is sensitive and easy to fall into local optimum solution, in order to optimize the initial cluster centers of K-means clustering algorithm. The chemotaxis of bacteria is simplified as the process of searching for the optimal solution of the particles in the particle swarm, and then the bacterias complete the further operation of reproduction and elimination-dispersal. The optimal solution of the hybrid algorithm is determined as the initial clustering center, and the disadvantages of the k-means algorithm are solved. The results on standard date set Iris, Wine, Glass of UCI display that the accuracy and stability of this proposed algorithm is higher than that of the popular clustering algorithms, at the same time, it can solve complex optimization problems most effectively.

Key words: k-means; bacterial foraging algorithm; particle swarm optimization; clustering algorithm

1 引言

K-means 算法原理简单, 运行速度快, 能够处理大量数据集, 因而成为了经典的无监督的、划分式聚类算法^[1]. 但由于对初始值敏感, 初始聚类中心不同会导致聚类结果不同, 容易陷入局部最优解, 稳定性较差, 进而影响聚类结果的准确性^[2]. 因此, 如何提

高聚类的准确率和稳定性, 研究人员在 k-means 算法的高效性上做了大量的研究.

近年来, 群智能算法和聚类算法的结合, 弥补了传统聚类算法的缺陷. 文献[3]将蚁群算法同聚类算法相结合, 蚁群算法的全局搜索可解决聚类算法易陷入局部最优的问题, 同时自动确定聚类数目避免聚类数目的不确定性, 聚类质量得到改善. 文献[4]

收稿日期: 2015-07-15; **修回日期:** 2016-09-10

基金项目: 国家高科技研究发展计划(“八六三”计划)(2014AA015204); 山西省国际科技合作项目(2014081018-2); 山西省科技基础条件平台建设基金项目(2013091003-0103)

将粒子群算法同 k-means 聚类算法相结合,把粒子群中的粒子做为聚类中心,粒子通过跟踪最优粒子的位置,更新自身位置,得到更好的聚类结果.文献[5]提出了细菌觅食与聚类相结合的算法,不断优化适应度函数,细菌最终的位置确定为聚类中心,仿真实验显示了算法具有较高的可行性和准确率.

K-means 算法初始聚类中心的随机选取直接影响聚类结果,本文提出了一种将 BFO-PSO 与 k-means 结合的聚类算法.

2 相关算法介绍

2.1 k-means 算法

传统的 k-means 算法随机选择空间中 n 个点作为聚类中心,将靠近 n 个点的数据进行归类^[6].通过迭代,逐次更新聚类中心的值,从而获得最优聚类结果.其中用欧式距离衡量数据对象与聚类中心的距离,向量 $\mathbf{x} = (x_1, x_2, \dots, x_n)$ 和向量 $\mathbf{y} = (y_1, y_2, \dots, y_n)$ 的欧式距离用 $d(x_i, y_i)$ 表示,定义如式(1):

$$d(x_i, y_i) = \left[\sum_{i=1}^n (x_i, y_i)^2 \right]^{1/2} \quad (1)$$

用准则函数评价聚类效果.目标对象为 o , o_i 表示聚类 C_i 的均值,准则函数定义如式(2):

$$E = \sum_{i=1}^k \sum_{o \in C_i} |o - o_i|^2 \quad (2)$$

假设数据集 Q 含有 n 个数据对象 $q_1 \dots q_n$, k 个聚类为 $w_1, w_2, \dots, w_k$,其具体步骤如下:

Step 1 从数据集 Q 中随机选取 k 个对象数据作为聚类算法的初始聚类中心 $a_1, a_2, \dots, a_k$

Step 2 利用欧式距离依据聚类中心对数据进行划分,如果欧式距离 $d(q_i, a_j) < d(q_i, a_m)$ 则将数据对象 q_i 划分到最近的聚类中心 a_i^*

Step 3 重新计算聚类中心,依据公式 $a_i = \frac{1}{n_i} \sum_{q \in W_i} q$ 得到新的聚类中心 $a_i (1 \leq i \leq k)$

Step 4 如果 $a_i^* = a_i$,则聚类中心不再改变,算法结束.

算法流程如图 1 所示.

K-means 的优点是:结构简单,能够处理大量数据集,适用于多种数据类型.缺点是初始聚类中心的选择容易陷入算法的局部最优,聚类数 k 的不同会导致不同的聚类结果.

2.2 基于粒子群优化的细菌觅食算法 BFO-PSO

BFO-PSO 算法是在细菌觅食的基础上结合粒子群的优化算法^[7].Passino 根据大肠杆菌的觅食特

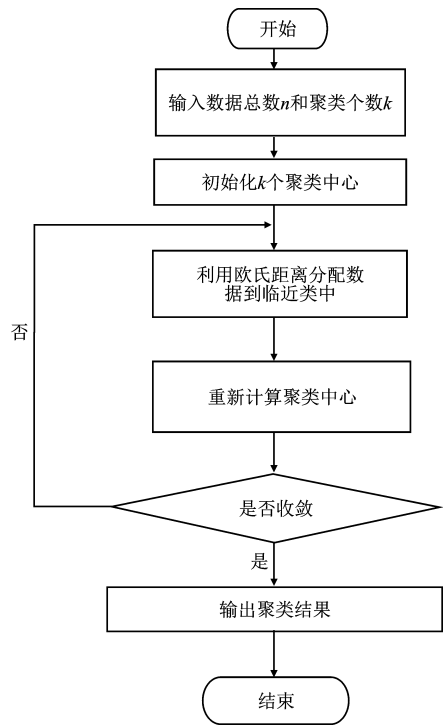


图 1 k-means 算法流程图

性提出了细菌觅食算法,包含趋向、复制、迁徙三个步骤^[8].趋向性是 BFO 算法的核心操作,包括细菌的旋转和游动.旋转的更新为式(3):

$$q_k(j+1, n, l) = q_k(j, n, l) + C(k) \cdot \Phi(k) \quad (3)$$

式中, $q_k(j, n, l)$ 用来表示个体 k 在经过 j 次趋向, n 次复制, l 次迁徙时的位置. $C(k)$ 表示步长, $\Phi(k)$ 表示旋转方向.细菌向任意方向移动单位步长定义为旋转,完成每次旋转检查适应度值,若改善则沿着同一方向游动,如此循环直到达到移动步数的上限或适应度值保持不变,此过程为游动.

复制操作根据适应度值淘汰搜索能力较差的半数细菌,剩下的半数细菌进行自我复制保持细菌总数不变.最后,细菌以一定概率进行迁徙,得到最优解.

趋向性使细菌获得较强的局部搜索能力,但细菌趋向的随机性导致收敛速度慢,而全局搜索性能较优的粒子群算法后期由于缺乏有效的机制,易陷入局部最优,因此 BFO-PSO 算法取长补短将全局搜索能力强的粒子群算法与局部搜索能力强的细菌觅食算法结合,完善了 BFO 全局搜索性,提高了 BFO 的优化能力,获得最优解.

将混合算法中粒子群中的粒子移动更新位置取代细菌觅食中利用细菌趋向操作更新位置,将趋化行为中的旋转和游动简化为一步,利用式(4)进行细

菌位置的更新,提高了细菌的收敛速度,用式(5)更新细菌的速度,达到迭代后次数进行 BFO 中的复制和迁徙操作。

$$Q_k(j+1, n, l) = q_k(j, n, l) + V_k(n+1) \quad (4)$$

$$V_k(n) = \omega_k^V(n-1) + a_2 \text{rand}_2(n-1) \\ (g_{k\text{best}}(n-1) - X_k(n-1)) \quad (5)$$

3 基于 BFO-PSO 的 k-means 算法模型

鉴于 BFO-PSO 算法的特点,本文将粒子群和菌觅食算法结合的算法用于改进 k-means 算法,提高聚类质量。融合粒子群算法的细菌觅食算法,首先利用粒子群算法进行全局搜索,将粒子做为细菌进行 BFO 的三大操作的趋向操作,通过复制操作进行优胜劣汰,幸存的细菌可作为 k-means 聚类的中心,而迁徙操作使得算法跳出局部最优,实现全局搜索。新的算法既保证了全局搜索性能又避免 k-means 随机选择聚类中心的弊端,过程如下所示。

Step 1 载入相关数据、初始化参数。细菌种群规模为 S , 趋向次数 N_c , 复制次数 N_{re} , 迁徙次数 N_{ad} , 游动次数 N_s , 迁徙概率 P_{ad} 。根据聚类数目 x , 任意选择 x 个对象作为细菌初始位置和初始聚类中心。

Step 2 对细菌的适应值进行评价。

Step 3 运行混合 BFO-PSO 算法三层循环。

(1)内层循环(趋化过程):粒子群中的粒子当做细菌通过式(4)更新位置,每次迭代更新粒子局部最优和全局最优,通过式(5)更新速度。满足迭代次数进入中层循环即复制,否则趋化次数+1 继续粒子的寻优。

(2)中层循环:复制操作。同时对速度进行更新。复制适应度值累加和排名在前的半数细菌。达到迭代次数后进行迁徙操作,否则复制次数+1 返回内层循环(趋化过程)。

(3)外层循环:迁徙操作。每个细菌生成一个随机数 v , 如果 $v < \text{迁徙发生概率 } p_{ad}$, 则该细菌被淘汰,重新随机生成一个新的个体。迁徙达到迭代次数则输出最优解,否则迁徙次数+1 返回内层循环(趋化过程)。

Step 4 输出最优解。

Step 5 运行聚类操作。把混合 BFO-PSO 的最优解作为 k-means 算法的聚类中心,运行 k-means 算法。

算法流程如图 2 所示。

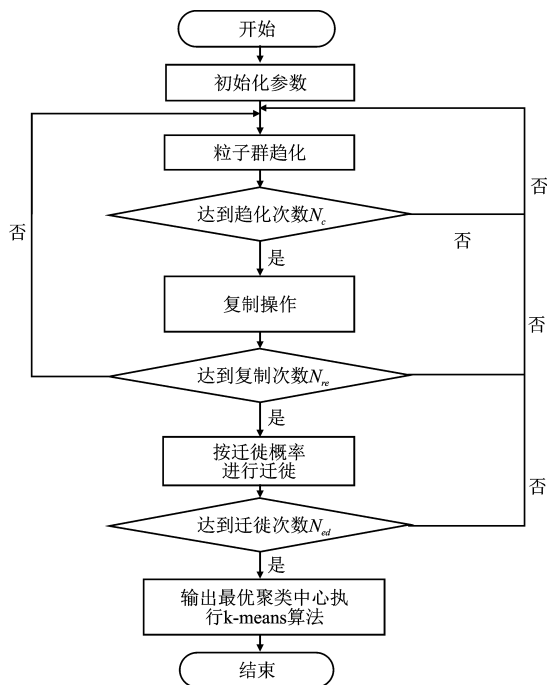


图 2 K-BFOPSO 算法流程图

4 实验结果与分析

4.1 主要参数设置

根据参考文献[9]算法常用推荐值和实验结果比较,参数如下选择时聚类结果较好。细菌种群规模为 $S=50$, 趋向次数 $N_c=100$, 复制次数 $N_{re}=4$, 迁徙次数 $N_{ad}=2$, 游动次数 $N_s=40$, 迁徙概率 $P_{ad}=0.25$, 步长 $C(k)=2$, $a_1=a_2=2$, $\omega=0.85$ 选择 windows 7 作为测试平台, Matlab 7.0 作为实验环境。

4.2 算法测试

选择 UCI 中的数据集 Iris、Wine、Glass 进行测试, Iris 是以鸢尾花为特征的 UCI 数据集, 包含 150 个实例, 每个实例包含 4 个属性, 分为 3 类, 每类含有 50 个数据。Wine 是以葡萄酒的成分分析为特征的数据集, 包含 178 个实例, 每个实例包含 13 个属性, 分为三类, 每类的数据数量为 48、59、71。Glass 是以玻璃材质为特征的数据集, 包含 214 个实例, 每个实例包含 9 个属性, 分为数量不等的六类。

实验前首先采用极大值标准化方法对高维数据进行降维处理, 避免数据各维取值范围不一致对聚类结果的影响^[10], 如式(6)所示:

$$x_{ij} = \frac{x_{ij}}{\max\{x_{ij}\}} \quad (i = 1, 2, \dots, m; j = 1, 2, \dots, n) \quad (6)$$

将本文算法 K-BFOPSO 和相关聚类算法运行 50 次, 单次最大迭代次数 100 次, 取各算法准确率的

平均结果,如表 1 所示. 其中聚类正确率公式:准确率 = $q/Q \times 100\%$. 其中 q 表示分配到某个类中的正确的数据个数, Q 表示该类中所包含的数据总数.

表 1 运行 50 次实验结果比较

数据集	聚类算法	准确率/%
iris	K-means	84.67
	PSO	89.01
	ACO	81.63
	K-BFO	91.34
	K-BFOPSO	94.96
wine	K-means	68.91
	PSO	70.32
	ACO	67.62
	K-BFO	91.34
	K-BFOPSO	80.11
Glass	K-means	67.83
	PSO	50.65
	ACO	60.11
	K-BFO	73.03
	K-BFOPSO	76.59

实验结果可知,本文算法 K-BFOPSO 与传统的 k-means 和常用的聚类算法 ACO、PSO,以及基于 BFO 的 K-BFO 算法相比,K-BFOPSO 算法明显提升了聚类的正确率,做为一种改进的 k-means 算法, k-means 的聚类效果得到改善.

为了更好的验证本文算法的高效性,图 3~图 5 为在实验数据集上分别运行 10 次、30 次、50 次、100 次的准确率平均结果. 不难看出作为聚类算法,本文提出的 K-BFOPSO 针对不同数据集运行不同的次数,当准确率用折线表示时,整体折线走势均高于其他聚类算法,并逐渐趋于平稳. 因此,K-BFOPSO 可作为一种基于全局最优的聚类算法.

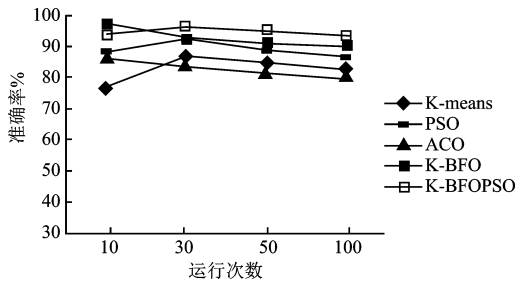


图 3 Iris 数据集各算法准确度对比

5 结束语

本文提出了一种基于 BFO-PSO 的改进聚类 k-

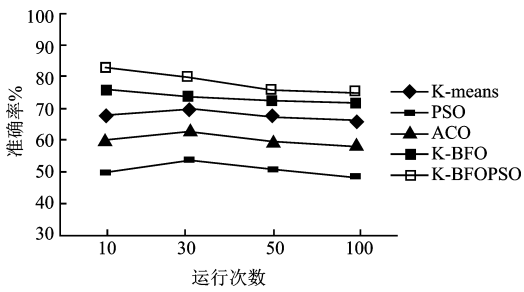


图 4 Glass 数据集各算法准确度对比

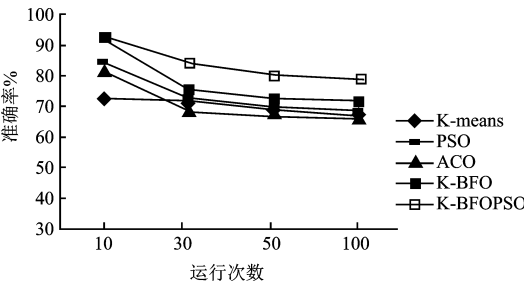


图 5 Wine 数据集各算法准确度对比

means 算法. 该算法将 PSO 嵌入 BFO,结合 BFO 良好的局部搜索能力和 PSO 全局搜索特性,形成 BFO-PSO 算法通过优化目标函数来找到 k-means 聚类算法的初始聚类中心. 理论分析和实验结果表明该算法明显提高了聚类准确率,是一种高效的聚类方法. 未来将对聚类时间方面做改进,降低该算法的聚类时间.

参考文献:

[1] 胡伟. 改进的层次 K 均值聚类算法[J]. Computer Engineering and Applications, 2013, 49(2):157-159.

[2] 曹永春,蔡正琦,邵亚斌. 基于 K-means 的改进人工蜂群聚类算法[J]. 计算机应用, 2014, 34(1): 204-207.

[3] 周新华,黄道. 一种基于蚁群算法的模糊 C 均值聚类[J]. 控制工程, 2005, 12(2): 132-134.

[4] 范会联,罗跃国,李献礼. 一种带克隆选择的粒子群动态聚类算法[J]. 山东大学学报(工学版), 2011, 41(2):12-17.

[5] Wan M, Li L, Xiao J, et al. Data clustering using bacterial foraging optimization[J]. Journal of Intelligent Information Systems, 2012, 38(2): 321-341.

[6] 叶志伟,尹宇洁,王明威,等. 一种基于杜鹃搜索算法的聚类分析方法[J]. 微电子学与计算机, 2015, 32(5): 105-110.

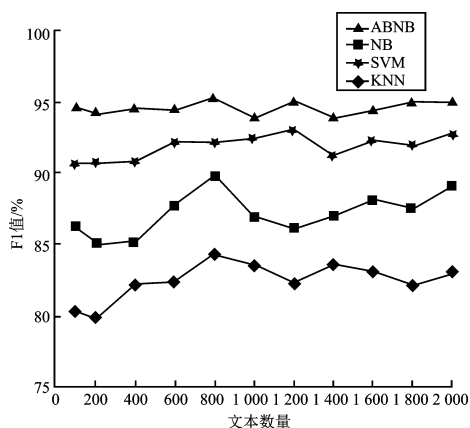


图3 算法 F1 值对比

于朴素贝叶斯的高效分类算法结合自适应提升算法,保证了 F1 值的平稳.实验结论表明分类器的设计合理,有优良的分类性能.可以广泛地应用在文章主题分析、垃圾邮件过滤、政府文档归类等应用上.本文有些地方还需要更做进一步的研究,比如机器学习技术在分词库建立的应用,多分类系统的设计,还有对海量数据进行处理技术.这些将会在后续内容中进一步探讨.

参考文献:

- [1] 靳小波. 文本分类综述[J]. 自动化博览, 2006(增刊 1):24-29.
- [2] 苏金树, 张博锋, 徐昕. 基于机器学习的文本分类技术研究进展[J]. 软件学报, 2006, 17(9):1848-1859.

- [3] 芦苇, 彭雅. 几种常用文本分类算法比较与分析[J]. 湖南大学学报: 自然科学版, 2007, 34(6):67-69.
- [4] 彭瑜. 基于语法的分词系统的设计与实现[D]. 成都: 电子科技大学, 2013.
- [5] Eyheramendy S, Lewis D, Madigan D. On the naive bayes model for text categorization[C]// The 9th International Workshop on Artificial Intelligence and Statistics. USA, Key West, 2003
- [6] 于玲, 吴铁军. 集成学习:Boosting 算法综述[J]. 模式识别与人工智能, 2004, 17(1):52-59.
- [7] 李航. 统计学习方法[M]. 北京:清华大学出版社, 2012: 75-90.
- [8] 李湘东, 曹环, 黄莉. 文本分类中训练集相关数量指标的影响研究[J]. 计算机应用研究, 2014, 30(11): 3324-3327.
- [9] 陈飞. 基于条件随机场方法的开放领域新词发现[J]. 软件学报, 2013, 24(5): 1051-1060.
- [10] 杨长春. 基于文本相似度的微博网络水军发现算法[J]. 微电子学与计算机, 2014, 31(3): 82-85.

作者简介:

徐凯 男, (1987-), 硕士研究生. 研究方向为推荐系统、数据挖掘. E-mail:504087493@qq.com.

陈平华 男, (1967-), 教授. 研究方向为云计算、Web 挖掘、推荐系统.

刘双印 男, (1977-5), 博士, 教授. 研究方向为智能计算、智能信息处理、物联网.

(上接第 62 页)

- [7] 毛力, 樊养余, 王慧琴, 等. 基于 PSO-BF 优化算法的关系数据库水印算法[J]. 计算机应用研究, 2014, 31(5): 1484-1487.
- [8] Shen H, Zhang M. Bacterial foraging optimization algorithm with quorum sensing mechanism[C]// Applied Mechanics and Materials. Singapore, 2014 (556): 3844-3848.
- [9] Wan M, Li L X, Xiao J H, et al. Data clustering using bacterial foraging optimization[J]. Journal of Intelligent Information Systems, 2012, 38(2):321-341.

- [10] 刘丽轻, 丁巧林, 张铁峰, 等. 数据预处理方法对模糊 C 均值聚类的影响[J]. 电力科学与工程, 2011, 26(8): 24-27.

作者简介:

闫婷 女, (1990-), 硕士研究生. 研究方向为人工智能、医学信息处理.

谢红薇(通讯作者) 女, (1962-), 博士, 教授. 研究方向为人工智能、医学信息学. E-mail: xiehongwei@tyut.edu.cn.