第 27 卷第 5 期　　四川理工学院学报(自然科学版)　　Vol. 27　No. 5
2014 年 10 月　　Journal of Sichuan University of Science & Engineering(Natural Science Edition)　　Oct. 2014
文章编号:1673-1549(2014)05-0064-04　　DOI:10.11863/j.suse.2014.05.15

基于 K-means 的改进差分进化聚类算法

乔艳霞, 邹书蓉, 张洪伟

(成都信息工程学院计算机学院, 成都 610225)

摘　要:K－means 聚类算法简单,收敛速度快,但是聚类算法的结果很容易受到初始聚类种群的影响,往往导致局部最优。差分进化算法具有很强的全局收敛能力和鲁棒性,但其收敛速度较慢。为此,将 K－means 聚类算法和差分进化算法相结合,提出一种基于 K－means 的改进差分进化聚类算法。该算法设置在一定范围内随迭代次数动态增加的交叉算子,以使算法在迭代过程中先进行全局搜索,再进行局部搜索,这样有助于平衡算法的全局寻优和局部搜索能力,并且加快了算法的收敛速度。最后,通过实验测试了算法的有效性。

关键词:差分进化;聚类;K－means;动态交叉算子

中图分类号:TP301.6　　**文献标志码**:A

引　言

K－means[1]是基于划分的经典聚类分析算法之一,算法简单、收敛速度快,但是聚类算法的结果很容易受到初始聚类种群的影响,往往导致局部最优。为解决这一问题,近年来许多学者应用各种智能算法优化 K－means聚类算法[2],主要是基于全局最优思想,如遗传算法、粒子群优化算法等,并取得了比较好的成效。

差分进化算法(differential evolution,DE)是 Storn 和 Price[3]在 1995 年提出的,它是一种基于种群差异和随机搜索的进化算法。DE 是基于仿生智能演化的计算算法,有内部信息共享和保留个体最优解的特征。差分进化计算特有的变异操作和良好的全局优化能力,能够使用个体信息和全局信息,指导搜索及优化是算法与其他算法最主要的区别。文献[4-5]将差分进化和K－means 算法相结合,与其他算法相比,DE－kmeans 具有能更好的搜索性能。但差分进化算法也存在一些缺点,如弱局部搜索能力,以及其搜索能力针对参数的特性有很大的依赖性等。许多学者从不同方面对 DE 进行改进[6-8],取得了很好的效果。

本文提出一种基于 K－means 的改进差分进化算法,其中设置动态的交叉概率算子,使交叉算子在一定范围内随着种群迭代次数的增加而增大,以使算法先进行全局搜索,再进行局部搜索,这样有助于二者能力的协调和平衡。实验结果表明,该算法不仅加快了收敛的速度,并具有良好的全局优化能力。

1　K－means 聚类算法

K－means 均值聚类的思想是:基于数据对象与集合(聚类)中心的距离,数据对象被分到与其距离最近(即该类中的数据样本之间最接近)的集合(聚类)[9]。K－means 算法的参数是 k,把含有 m 个对象的集合划分到与之最近的子集中,并且同一子集内对象之间比较相似,而不同子集之间数据对象差异很大。

该算法的基本步骤[10]如下:

步骤 1:从样本集中选取 k 条数据样本当作初始的 k

收稿日期:2014-04-13

作者简介:乔艳霞(1987-),女, 河南周口人,硕士生,主要从事智能工程方面的研究,(E-mail)qiaoyanxia2012@163.com

个集合(类簇)的中心;

步骤 2:根据最近邻法则,数据样本被分到与之最接近的类簇;

步骤 3:计算新的类簇中心,它是由计算每个集合中所有样本数据的均值得到的;

步骤 4:类簇中心不再改变作为算法终止标准,返回最终的结果。否则,返回步骤 2 继续计算。

其核心问题是:如何选择算法的初始类簇中心。如果初始划分和全局最优的划分严重偏离,该算法可能会趋向于得到局部的最优值。

2　标准差分进化算法

标准差分进化算法的核心操作包含变异操作、交叉操作以及选择操作三个部分[3]。

变异操作:变异操作产生的新个体是由群体内多个单独个体的线性运算(差分运算)得到的。DE 包括多种变异机制,其中采用最多的是由式(1)表示的变异策略。

$$v_i(g) = c_{r_1}(g) + F \cdot [c_{r_2}(g) - c_{r_3}(g)] \tag{1}$$

其中,NP 为种群中的染色体数目,F 为变异算子,$c_{r_1}(g)$ 为第 g 代种群中的一个个体,$i = 1,2,3\cdots NP, r_1, r_2, r_3 \in [1, NP]$,且 $r_1 \neq r_2 \neq r_3 \neq i$。

交叉操作:通过将当前个体 c_i 的部分分量用目标个体 v_i 的对应分量替换,从而生成测试个体 t_i。常见的交叉方式有两种:指数交叉和二项交叉。其中二项交叉具体执行方式为:首先对每条染色体的每一个分量都生成一个 0 ~ 1 之间的随机数 r,若 $r < CR$,则接受目标个体的对应分量,否则保留当前个体的对应分量。即:

$$t_{i,j}(g) = \begin{cases} v_{i,j}(g), r < CR \text{ 或 } j = rndm(\{1,2,\cdots,n\}) \\ c_{i,j}(g), \text{其他} \end{cases} \tag{2}$$

其中,CR 是交叉算子,n 为染色体的属性个数。

选择操作:标准 DE 采用贪心选择,当前个体 c_i 与测试个体 t_i 相比,选择更好的个体进入下一代搜索。即:

$$c_i(g+1) = \begin{cases} t_i(g), f(t_i(g)) < f(c_i(g)) \\ c_i(g), \text{其他} \end{cases} \tag{3}$$

其中,$f(t_i(g))$ 为适应度函数,即目标函数。

3　动态递增的交叉算子

交叉算子系数越大,局部搜索更好,相反,交叉因子越小,则会有更强的全局搜索能力[11]。为了平衡算法的全局寻优能力和局部搜索能力,本文采用在一定范围内,跟随迭代次数动态变化的交叉算子:

$$CR = CR_{\max} - (CR_{\max} - CR_{\min}) * \exp(1 - g_{\max}/(g_{\max} + 1 - g)) \tag{4}$$

其中 $(CR_{\min}, CR_{\max})$ 是 CR 的上下限。

4　算法步骤

(1) 初始化

初始化算法的各个参数:如种群中染色体的条数 NP,簇数 K,(变异)缩放因子 F,交叉算子 CR 的最大值 $CR_{\max}$ 和最小值 $CR_{\min}$,最大迭代次数 $g_{\max}$,当前迭代数 g。

(2) 初始化种群

对有 N 个样本组成的数据集 $U = \{x_1\ x_2\ x_3 \cdots x_n\}$ 里的数据,选择 k 条用作初始种群。用 K – means 聚类算法对数据样本进行分类,根据最近邻法则,把数据样本划分为 k 个的子集。分类的标准则依据欧几里得公式:

$$distance(O_i, O_j) = \|O_i - O_j\| = \sqrt{|O_{i1} - O_{j1}|^2 + |O_{i2} - O_{j2}|^2 + \cdots + |O_{in} - O_{jn}|^2} \tag{5}$$

其中 $(O_{i1}, O_{i2}, O_{i3}, \cdots, O_{in})$ 为第 i 条数据样本的 n 个属性。

(3) 计算适应度函数

目标函数是衡量种群染色体之间距离的标准,即类簇中心之间的距离,函数值越小表示两个类簇之间距离越大。目标函数是为每个个体的价值进行计算和评估,并保留最优函数值。通常采用的目标函数为:

$$f(e) = \sum_{k=1}^{K} \sum_{i=1}^{m_k} \|x_i^k - c_k\| \tag{6}$$

其中 $f(e)$ 表示同属于一个类簇的所有样本的离散度,它越小,则表示该类内的数据越相似(即样本点越集中)。

(4) 变异操作

对种群中的每一条染色体,按照公式(1)执行变异操作,生成目标个体,其中公式中的 3 个个体是随机选择的。

(5) 交叉操作

按照公式(4)计算交叉算子,通过将当前个体的部分分量用目标个体的对应分量替换,从而生成测试个体。

(6) 选择操作

应用公式(3)选择测试的个体和当前的个体中较好的,保留最佳个体的目标函数值。

(7) 判断是否结束

确定该算法是否满足终止条件(算法收敛到最优解),或确定是否差分进化算法来达到最大迭代次数 g_{max}。如果没有满足条件,则转向步骤 3 继续执行;如果达到最大迭代次数 g_{max},算法结束,输出结果。

5 实验仿真及结果

实验环境是:2.99 G CPU,1.97 G 内存。visual C++6.0,C 语言编程。在 UCI 中选择三个数据集作为测试数据集,样本集的特性见表 1。参数设置如下:种群大小为数据集样本属性个数的 10 倍,g_{max} 为 50,其中 F 为 0.5,CR 的最大值 CR_{max} 为 0.9,最小值 CR_{min} 为 0.1。DE – *km*eans[2-3] 算法 F 为 0.5,CR 为 0.1。每种算法分别运行 50 次,实验结果见表 2。

表 1 样本集的特性

样本集	样本数目	属性个数	类数
iris	150	4	3
wine	178	13	3
Zoo	101	16	7

表 2 算法的比较结果

样本集	使用的算法	最小目标函数值	最大目标函数值	目标函数值的平均	时间(s)
Iris	K – means 算法	97.0435	119.8022	108.9374	0.06
	DE – *km*eans	97.1019	97.6623	97.3524	4.19
	文献[12]的算法	97.1789	97.2425	97.1765	2.89
	本文算法	96.9309	97.2322	97.2134	1.36
Zoo	K – means 算法	117.5623	123.9463	119.4236	0.09
	DE – *km*eans	111.7325	120.9653	117.6542	29.51
	文献[12]的算法	108.7648	113.4453	109.5682	17.14
	本文算法	108.0549	116.7129	112.7356	11.56
Wine	K – means 算法	1.6603E+04	1.6815E+04	1.6924E+04	0.18
	DE – *km*eans	1.6546E+04	1.6581E+04	1.6553E+04	17.35
	文献[12]的算法	1.6517E+04	1.6528E+04	1.6524E+04	9.45
	本文算法	1.6503E+04	1.6557E+04	1.6522E+04	5.93

从表 2 可以看出,K – means 算法具有快速收敛的速度,但优化准确性差。DE – *km*eans 算法优化精度还是不错的,但收敛速度慢。与上述两种方法相比,改进的算法的收敛的速度及优化准确性更好,优化结果也更稳定。

6 结束语

本文在基本差分进化算法的基础上提出动态交叉算子,将其和 K – means 相结合,以提高收敛速度。实验结果表明,该算法收敛速度快,既具有差分进化算法的全局优化能力,又保留了 K – means 聚类算法的优点,具有快速搜索速度和稳定的结果。

参 考 文 献:

[1] MacQueen J. Some methods for classification and analysis of multivariate observations[C]//Proc.of the 5th Berkeley Symposium on Mathematics Statistic Problem, Berkeley, june 21-July 18,1967:281-297.

[2] 王雪梅,李晓峰,高巍巍.一种改进的 K-means 聚类算法研究[J].计算机与数字工程,2013,41(11):1717-1719,1759.

[3] Storn R,Price K.Differential Evolution:A simple and efficient heuristic for global optimization over continuous spaces[J].Journal of Global Optimization,1997(11):341-359.

[4] Paterlini S,Krink T.High performance clustering with differential evolution[C]//Proc.of Evolutionary Computation, 2004,California, june 19-23,2004:2004-2011.

[5] Sudhakar G.Effective image clustering with differential evolution technique[J].International Journal of Computer

and Communication Technology,2010,2(1):11-19.

[6] 董明刚,王 宁,程小辉.改进的组合差分进化优化算法[J].计算机仿真,2013,30(1):389-392.

[7] 许奕昕,白 焰,赵天阳,等.泊松分布下无线传感器网络多目标覆盖控制[J].计算机应用,2013,33(7):1820-1824,1832.

[8] 姜立强,强洪夫.带基向量种群的改进差分进化算法[J].计算机工程,2012,38(3):9-11.

[9] 毛韶阳,林肯立.优化 K-means 初始聚类中心研究[J].计算机工程与应用,2007,43(22):179-181.

[10] 孙吉贵,刘 杰,赵连宇.聚类算法研究[J].软件学报,2008,19(1):48-61.

[11] 邓哲喜,操敦虔,刘晓冀,等.一种新的差分进化算法[J].计算机工程与应用,2008,44(24):40-42.

[12] 高 平,毛 力,宋宜春.基于改进差分进化的 K-均值聚类算法[J].电脑知识与技术,2013,9(22):5064-5067.

Modified Differential Evolution Clustering Algorithm Based on K-means

QIAO Yanxia, *ZOU Shurong*, *ZHANG Hongwei*

(College of Computer Science & Technology, Chengdu University of Information Technology, Chengdu 610225, China)

Abstract: K-means clustering algorithm is simple and converge rapid, but the result of clustering algorithm is vulnerable to the influence of initial cluster population, which often leads to local optimum. Differential evolution algorithm has strong global convergence ability and robustness, but it's convergence volocity is slow. For this reason, K-means clustering algorithm is combined with the differential evolution algorithm, then a modified differential evolution clustering algorithm based on K-means is proposed. Within a certain range of the algorithm, the crossover operators that increased dynamically with iterative number are set, so that the algorithm carries out global search first and local search second in the iterative process, which can help to balance the global search capability and local search capability of algorithm, and accelerate the convergence speed of algorithm. Finally, the experiments have tested the effectiveness of algorithm.

Key words: differential evolution; clustering; K-means; dynamic crossover operator