

基于差分进化算法的 K -Means 算法改进

刘莉莉, 曹宝香

(曲阜师范大学 信息科学与工程学院, 山东 日照 276826)

摘要:针对现如今传统的 K -Means 聚类算法所普遍存在的对初始聚类中心选择敏感且易陷入局部最优解的问题,文中将全局寻优能力较强的差分进化算法引入该算法中,其中通过采用选择结构的多模式进化方案、自适应调整的控制参数,从而提出了一种性能优良的改进的差分进化算法。同时进一步将改进的差分进化算法和 K -Means 聚类算法相结合,得以较好地解决了 K -Means 聚类算法中初始聚类中心的优化问题。通过在三种国际通用数据集上进行实验测试,最终的实验结果表明,该方法可以明显加快算法收敛速度,增强全局优化能力,并且有效提高了聚类结果的质量和稳定性。

关键词:聚类算法; K -Means 聚类算法; 差分进化算法; 进化模式; 控制参数

中图分类号: TP399

文献标识码: A

文章编号: 1673-629X(2015)10-0088-05

doi: 10.3969/j.issn.1673-629X.2015.10.019

Improvement of K -Means Algorithm Based on Differential Evolution Algorithm

LIU Li-li, CAO Bao-xiang

(College of Information Science and Engineering, Qufu Normal University, Rizhao 276826, China)

Abstract: According to the defects of traditional K -Means clustering algorithm such as sensitive to the initial clustering center selection, falling into the local optimal solution easily, the differential evolution algorithm which has the rich global search ability is introduced in this paper, then an improved differential evolution algorithm with multi-model evolution scheme of selection structure and control parameters of adaptive adjustment is presented in the meantime. The algorithm combined with K -Means algorithm has solved initial center optimization problem well. Experiments on the international datasets show that this method could speed up the convergence speed significantly, enhancing the ability of global optimization, improving the clustering quality and stability effectively.

Key words: clustering algorithm; K -Means clustering algorithm; differential evolution algorithm; evolution pattern; control parameter

1 概述

K -Means 算法, 又被称为 GLA (Generalized Lloyd Algorithm) 算法, 它是由 J. B. MacQueen 于 1967 年提出的一种经典的基于划分的聚类算法^[1]。作为一种简单的迭代型聚类算法, 它以 Lloyd 迭代法分割数据集的方式将一个给定的数据集划分为用户指定的 K 个簇, 具有算法简单高效、收敛速度快、便于处理大型数据集等优点, 现如今已经被广泛应用于科学研究和工业应用等诸多领域。但是, 传统的 K -Means 算法还具有以下缺点: 聚类个数 K 通常是用户依据经验事先给定的, 在初始化聚类中心时, 初始点也是任意选取的。所以具有随机性和不确定性, 使得聚类结果不稳

定且易陷入局部最优解。因此, 需要提出一种能改进 K -Means 算法以上缺点的全局优化算法。

差分进化算法 (Differential Evolution, DE) 是由 Storn R 和 Price K 于 1995 年为了解决切氏多项式问题而提出的一种新的进化算法^[2-4]。作为一种随机的并行全局搜索算法, 它采用实数矢量编码在连续空间中进行随机搜索。差分进化算法主要有以下优点: 结构简单, 控制参数少, 易于实现; 具有良好的鲁棒性和强大的全局寻优能力。通过基于变异、交叉、选择的差分算法对多个个体组成的种群进行操作, 使个体一代一代得以优化, 逐步逼近最优解, 实现种群的优化。

目前很多学者已经将遗传算法、蚁群算法、微粒群

收稿日期: 2014-11-22

修回日期: 2015-03-12

网络出版时间: 2015-09-23

基金项目: 山东省自然科学基金项目 (ZR2009GM009); 山东省科技攻关项目 (2012GGB01193)

作者简介: 刘莉莉 (1990-), 女, 硕士研究生, 研究方向为企业信息化与系统集成、云计算; 曹宝香, 教授, 硕士研究生导师, CCF 高级会员, 研究方向为企业信息化与系统集成、云计算。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20150923.1506.060.html>

算法等全局优化算法引入 K -Means 聚类算法中^[5-6],有效优化了聚类中心,取得了不错的效果。与以上优化算法相比,差分进化算法的优势更加明显,用变异、交叉、选择操作来取代传统 K -Means 聚类算法中不断更新聚类中心的过程,既可以有效避免 K -Means 算法陷入局部最优,取得高质量的初始聚类中心,还加快了差分进化算法的收敛速度^[7-8]。在此基础上,文中提出了一种改进的差分进化算法,实现了将改进的差分进化算法和 K -Means 聚类算法相结合,得以较好地解决了 K -Means 聚类算法初始中心的优化问题,有效提高了聚类质量和收敛速度。最后通过实验与 K -Means 聚类算法和传统的基于差分进化算法的 K -Means 算法相比较,聚类结果得到了有效的改进。

2 算法设计

2.1 传统的 K -Means 聚类算法

K -Means 聚类算法是一种典型的基于距离的划分聚类算法,通常采用误差平方和函数作为优化的目标聚类准则函数。基本思想是首先从含有 n 个数据对象的数据集中随机选择 K 个数据对象作为初始中心,然后计算每个数据对象到各中心的距离,根据最近邻原则,所有数据对象将会被划分到离它最近的那个中心所代表的簇中,随后分别计算新生成的各簇中数据对象的均值作为各簇新的中心,比较新的中心和上一次得到的中心,如果新的中心没有发生变化,则算法收敛,输出结果,如果新的中心和上一次的中心相比发生变化,则要根据新的中心对所有数据对象重新进行划分,直到满足算法的收敛条件为止^[9]。具体算法流程如下:

输入:簇的数目 K 值及含有 n 个数据对象的数据集 X 。

输出:使误差平方和 E 达到最小的 K 个簇。

(1) 在含有 n 个数据对象的数据集 X 中随机选择 K 个数据对象作为初始聚类中心;

(2) 分别计算数据集中每个数据对象到各个聚类中心的距离,根据最近邻原则将数据对象逐个划分到离其最近的聚类中心所代表的簇中,计算误差平方和准则函数的值;

(3) 分别计算各个簇中所有数据对象的均值作为各个簇的新的中心,以新的聚类中心来计算误差平方和准则函数 E 的值;

(4) 将步骤(3)计算得到的值和步骤(2)计算得到的值进行比较,如果两者差值的绝对值不大于预先设定的阈值,即聚类准则函数收敛,则转步骤(5),否则转步骤(2);

(5) 输出 K 个簇。

K -均值聚类算法的主要流程如图 1 所示。

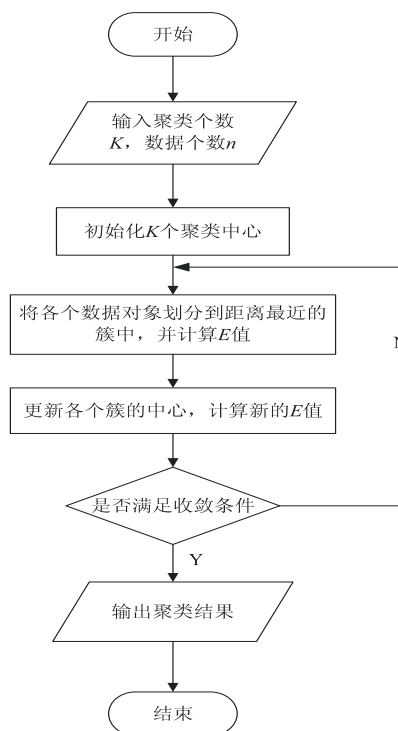


图 1 K -Means 聚类算法的基本流程

2.2 基于差分进化算法的 K -Means 聚类算法

该算法首先需要将从数据集中随机选出的聚类中心进行编码,构造初始种群,然后执行差分进化算法的变异、交叉、选择等操作,获得最佳个体,最后将最佳个体解码,得到最佳初始聚类中心并进行聚类。下面简单介绍基于差分进化算法的 K -Means 聚类算法的基本步骤。

输入:聚类数目 K 、含有 n 个数据对象的数据集 X 、种群规模 N 、缩放系数 F 、交叉概率 C_r 。

输出:输出最佳聚类结果。

(1) 编码:采用实数编码对从数据集中随机选出的聚类中心进行编码,一个编码对应一个可行解;每个个体是由 K 个聚类中心构成的向量串,由于样本向量维数为 D ,所以每个个体都是 $K \times D$ 维向量。具体编码方式如下:

$$X_i(t) = (c_{i,1}, c_{i,2}, \dots, c_{i,K}) \quad (i = 1, 2, \dots, N)$$

其中, $X_i(t)$ 表示第 t 代种群的第 i 个个体,迭代次数 t 的初始值为 0; $c_{i,j} (i = 1, 2, \dots, N; j = 1, 2, \dots, K)$ 表示第 i 个个体的第 j 个聚类中心。

(2) 种群初始化:随机从数据集中选出 K 个数据样本作为初始种群的一个个体,重复进行 N 次操作,构造出初始种群。

(3) 计算出个体 $X_i(t)$ 的适应度值 $f(X_i(t))$, 目的在于评价种群中的每个个体。设个体的适应度函数为: $f(X_i(t)) = 1/E$ 。

(4) 变异操作:变异操作基于当前种群中个体的

基因位进行。从当前种群中随机选择三个个体 $X_{a,j}(t), X_{b,j}(t), X_{c,j}(t)$, 且 $a \neq b \neq c \neq i$, 按下式计算得到变异个体 $V_i(t)$:

$$v_{i,j}(t) = x_{a,j}(t) + F(x_{b,j}(t) - x_{c,j}(t))$$

其中, F 为缩放系数。

(5)交叉操作: 变异个体 $V_i(t)$ 和种群中的个体 $X_i(t)$ 进行交叉操作, 按下式计算得到中间试验个体 $U_i(t)$:

$$u_{i,j}(t) = \begin{cases} v_{i,j}(t) & \text{if } \text{rand}(0,1) \leq C_R \text{ or } j = \text{rand}(i) \\ x_{i,j}(t) & \text{else} \end{cases}$$

其中, $\text{rand}(0,1)$ 是 $(0,1)$ 上服从均匀分布的随机数; C_R 为交叉概率, 且 $C_R \in [0,1]$; $\text{rand}(i)$ 是 $[1, K]$ 上的随机数。

(6)选择操作: 在当前进化个体 $X_i(t)$ 和中间试验个体 $U_i(t)$ 之间, 计算并比较适应度值, 采用贪心算法决定适应度函数值最小的个体, 即最佳个体进入下一代种群。

(7)对种群 $X(t+1)$ 中的个体进行检验, 如果满足终止条件, 则输出最佳个体, 算法终止; 否则将迭代次数 t 加 1, 返回第 2 步继续操作。

(8)将输出的最佳个体进行解码得到对应的最佳聚类中心集合, 根据最近邻原则将种群中的所有数据对象划分到相应的簇中。

(9)输出聚类结果。

该算法的流程图如图 2 所示。

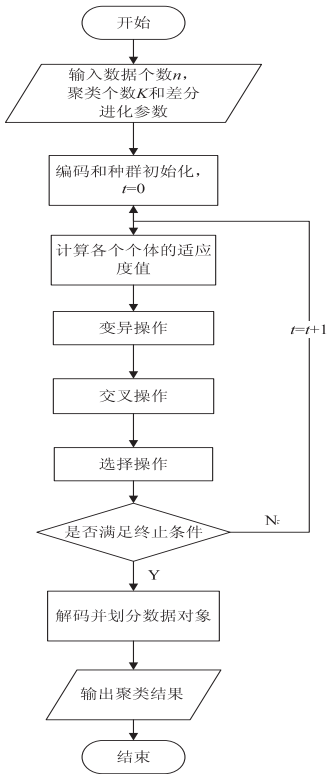


图 2 基于差分进化算法的 K -Means 聚类算法的基本流程

2.3 算法改进

将差分进化算法应用于 K -Means 聚类算法的初始聚类中心的优化, 提高了聚类质量, 并且算法设计简单, 容易实现, 通过变异和交叉操作保证了种群进化的多样性和算法的全局搜索能力^[7-8]。但是, 算法的局部搜索能力还有待加强, 尤其是算法进化后期的收敛速度也有待提高。

因此, 在不影响算法全局寻优能力的前提下, 文中提出了一种可以加强算法局部搜索能力的改进方案。

鉴于差分进化算法具有多种进化模式 (Evolution Strategy), 并且每种模式有其自身的性能特点, 因此考虑将这些进化模式进行有机结合来改进算法性能, 从而充分发挥各种进化模式的特点; 另外, 采用自适应的方式完善差分进化算法的相关参数, 使其在迭代过程中实现自适应调整, 降低算法对参数的敏感度。

差分进化算法的进化模式一般表示为: $DE/x/y/z$ ^[10]。其中, x 表示种群中的待变异个体是“随机产生的”或者是“当前种群中的最优个体”; y 表示参与交叉操作的差异向量的个数; z 表示交叉操作采用的方式, 包括“二项式交叉”和“指数交叉”。进化模式种类总结如表 1 所示。

表 1 差分进化模式总结

编号	差分进化模式	编号	差分进化模式
1	DE/best/1/bin	2	DE/best/1/exp
3	DE/rand/1/bin	4	DE/rand/1/exp
5	DE/rand-to-best/1/bin	6	DE/rand-to-best/1/exp
7	DE/best/2/bin	8	DE/best/2/exp
9	DE/rand/2/bin	10	DE/rand/2/exp

由于在进行交叉操作时, 采用二项式交叉的方式可以使子代个体与中间试验个体及父代个体交换更多信息, 相对于指数交叉方式更加有利于算法收敛, 因此通常选择二项式交叉作为交叉方式。

标准差分进化算法采用 $DE/rand/1/bin$ 进化模式, 它的特点是自由探索性突出, 全搜索能力较强, 不易陷入局部最优, 但收敛速度较慢。 $DE/best/1/bin$ 的特点是自由探索特性相对较弱, 局部搜索能力较强, 收敛速度较快, 但容易陷入局部最优。 $DE/rand-to-best/1/bin$ 模式的特点是自由探索性相对均衡, 具有良好的适应性, 缺点是鲁棒性较弱。

这三种进化模式的不同主要在于变异方式的差异, 从而导致不同的求解性能。为充分发挥以上三种进化模式的优点, 克服自身缺点, 现将这三种进化模式的变异方式采用选择结构结合起来, 算法其他部分与标准差分进化算法一致, 具体变异方式如下:

$$v_{i,j}(t)=\begin{cases}x_{a,j}(t)+F(x_{b,j}(t)-x_{c,j}(t)) & \text{if rand}(0,1)\leq C_R \text{ or } j=\text{rand}(i) \text{ and } M=1 \\ x_{\text{best},j}(t)+F(x_{a,j}(t)-x_{b,j}(t)) & \text{if rand}(0,1)\leq C_R \text{ or } j=\text{rand}(i) \text{ and } M=2 \\ x_{i,j}(t)+F(x_{\text{best},j}(t)-x_{i,j}(t))+F(x_{a,j}(t)-x_{b,j}(t)) & \text{if rand}(0,1)\leq C_R \text{ or } j=\text{rand}(i) \text{ and } M=3 \\ x_{i,j}(t) & \text{else}\end{cases}$$

差分进化算法主要有三个控制参数:种群规模 N 、缩放系数 F 、交叉概率 C_R 。尤其是 F 和 C_R 对算法性能有着极其重要的影响。 F 控制个体间差异的缩放程度, C_R 影响变异个体对中间试验个体的操作。在算法进化的前期阶段应保持种群个体的多样性,进行全局寻优,而在算法进化的后期阶段应加强局部搜索能力,加快算法的收敛速度。为了符合这种要求,在结合现有研究成果^[11-14]的基础上,文中提出采用取值随进化代数增加而逐渐递减的动态缩放系数,以及根据进化代数递增的动态交叉概率,从而可以很好地平衡算法的局部搜索能力和全局搜索能力。具体策略如下:

$$F=F_{\max}-(F_{\max}-F_{\min})\left(\frac{t}{T}\right)^2$$
$$C_R=C_{R_{\min}}+(C_{R_{\max}}-C_{R_{\min}})\left(\frac{t}{T}\right)^2$$

其中, F_{\max} 表示 F 的上限, $F_{\max}=0.9$; F_{\min} 表示 F 的下限, $F_{\min}=0.4$; $C_{R_{\max}}$ 表示 C_R 的上限, $C_{R_{\max}}=0.9$; $C_{R_{\min}}$ 表示 C_R 的下限, $C_{R_{\min}}=0.3$; t 表示当前进化代数; T 表示算法规定的最大进化代数。

按上式这样实现参数的自适应调整,可以保证随着进化代数的增加,缩放系数 F 逐渐增加,交叉概率 C_R 逐渐减小,进而保证算法在进化前期有较好的全局搜索能力,以及在后期有较快的收敛速度。

基于改进的差分进化算法的 K -Means 聚类算法步骤如下:

输入:聚类数目 K , 含有 n 个数据对象的数据集 X , 种群规模 N , 缩放系数 F , 交叉概率 C_R 。

输出:输出最佳聚类结果。

(1)对聚类中心进行编码,完成种群初始化;计算个体的适应度值以评价种群中的每个个体。

(2)对当前种群中的个体进行变异操作得到变异个体, F 根据上式进行动态递减策略取值。

(3)对当前种群中的个体进行交叉操作得到中间试验个体, C_R 根据上式进行动态递减策略取值。

(4)通过贪心算法选择适应度值最小的最佳个体进入下一代种群。

(5)对下一代种群中的个体进行检验,如果满足终止条件,则输出最佳个体,算法终止;否则将迭代次数加 1,返回第 2 步继续操作。

(6)将输出的最佳个体进行解码得到对应的最佳聚类中心集合,根据最近邻原则将种群中的所有数据对象划分到相应的簇中。

(7)输出聚类结果。

3 实验结果与分析

仿真实验的环境:操作系统为 Windows XP2,处理器为 AMD Athlon 5200+, 内存为 1 GB, 仿真软件为 Matlab7.0。采用国际上常用的 UCI 数据库中的 3 个数据集: Iris、Glass 及 Vowel 作为测试数据集。各数据集所含数据样本个数、数据样本的属性个数、类别个数如表 2 所示。分别将这 3 个数据集用于测试 K -Means 聚类算法、基于差分进化算法的 K -Means 聚类算法以及文中提出的基于改进差分进化的 K -Means 聚类算法的聚类效果。

表 2 测试数据集相关信息

数据集	数据样本个数	样本属性个数	类别个数
Iris	150	4	3
Glass	214	9	6
Vowel	990	10	11

在基于差分进化的 K -Means 聚类算法的仿真实验中,种群规模设置为种群个体属性维数的 10 倍,缩放系数 F 设置为 0.6,交叉概率 C_R 设置为 0.5,最大迭代次数 T 设置为 1 200;在基于改进的差分进化的 K -Means 聚类算法的仿真实验中,种群规模也取相应数据集样本属性维度的 10 倍,缩放系数 F 和交叉概率 C_R 按照公式自适应取值,最大迭代次数 T 也设置为 1 200。下面分别用 3 个测试数据集对 K -Means 聚类算法、基于差分进化的 K -Means 聚类算法以及文中提出的基于改进差分进化的 K -Means 聚类算法进行 50 次仿真实验。结果如表 3~6 所示。

表 3 在 Iris 数据集上的实验结果

	K -Means 聚类算法	基于差分进化的 K -Means 聚类算法	改进算法
最小类内距离	90.268 5	88.225 4	85.779 3
最大类内距离	185.457 4	98.657 8	91.558 4
平均类内距离	120.482 3	92.892 6	88.897 4

表 4 在 Glass 数据集上的实验结果

	K -Means 聚类算法	基于差分进化的 K -Means 聚类算法	改进算法
最小类内距离	20.290 6	19.248 5	18.897 6
最大类内距离	42.418 8	26.446 2	23.243 5
平均类内距离	31.354 7	22.054 2	20.756 2

表 5 在 Vowel 数据集上的实验结果

	$K - Means$ 聚类算法	基于差分进化的 $K - Means$ 聚类算法	改进算法
最小类内距离	5 984.587 3	5 622.938 7	5 021.862 7
最大类内距离	9 504.290 1	6 875.523 7	5 884.624 4
平均类内距离	7 638.317 8	6 174.237 3	5 183.749 2

表 6 在数据集上的算法平均收敛代数

	基于差分进化的 $K - Means$ 聚类算法	改进算法
Iris 数据集上算法 平均收敛代数	67	42
Glass 数据集上算法 平均收敛代数	685	445
Vowel 数据集上算法 平均收敛代数	1 097	856

通过表 3~5 可以看出,随机选择初始聚类中心的 $K - Means$ 聚类算法聚类结果波动范围较大,稳定性较差。相比之下,基于差分进化的 $K - Means$ 聚类算法和文中提出的改进算法的最小类内距离、最大类内距离,以及二者差值都明显缩小了。根据聚类的性质:类内距离越小,则同一簇中的数据对象之间越紧密,聚类质量越好;反之类内距离越大,同一簇内数据对象之间越松散,聚类质量越差。因此实验数据验证了基于差分进化的 $K - Means$ 聚类算法和改进算法对于提高聚类结果的稳定性和有效性有很大进步,显著改善了聚类质量。并且从结果分析可知,文中提出的改进算法的聚类结果和质量还要明显优于原基于标准差分进化算法的 $K - Means$ 聚类算法。

通过表 6 可以看出,文中提出的改进算法的平均收敛速度要明显快于基于差分进化的 $K - Means$ 聚类算法,从而证明了基于多模式进化方案和自适应控制参数的改进算法在提高算法收敛速度和优化全局寻优方面的有效性。

4 结束语

文中将具有结构简单、易于实现、控制参数较少、收敛速度较快等优点的差分进化算法引入 $K - Means$ 聚类算法中,并对算法做出如下改进:将多种进化模式进行有机结合,采用自适应的动态参数策略。实验结果表明,与传统 $K - Means$ 聚类算法相比较,文中提出

的改进算法对初始聚类中心的选择优化能力有显著提高,算法整体收敛速度较快,全局搜索能力更强,有效提高了聚类结果的稳定性,明显改善了聚类质量。

参考文献:

[1] MacQueen J. Somemethods for classification and analysis of multivariate observations[C]//Proc of the 5th Berkeley symposium on mathematical statistics and probability. Berkeley, America:University of California Press,1967.

[2] Storn R. Differential evolution design of an IIR-filter[C]//Proc of IEEE international conference on evolutionary computation. Nagoya, Japan:IEEE,1996.

[3] Storn R. On the usage of differential evolution for function optimization[C]//Biennial conference of the North American on fuzzy information processing society. Berkeley:IEEE,1996:519-523.

[4] Price K. Differential evolution;a fast and simple numerical optimizer[C]//Biennial conference of the North American on fuzzy information processing society. Berkeley:IEEE,1996:524-527.

[5] 赖玉霞,刘建平,杨国兴. 基于遗传算法的 K 均值聚类分析[J]. 计算机工程,2008,34(20):200-202.

[6] 李 飞,薛 彬,黄亚楼. 初始中心优化的 K-Means 聚类算法[J]. 计算机科学,2002,29(7):94-96.

[7] 刘凤龙,陈 曦,曹 敦. 基于差分演化的 K-均值聚类算法[J]. 计算技术与自动化,2010,29(1):48-50.

[8] Babu B V, Jehan M M L. Differential evolution for multiple-objective optimization[J]. Evolutionary Computation,2003,11(4):8-12.

[9] 欧陈委. K-均值聚类算法的研究与改进[D]. 长沙:长沙理工大学,2011.

[10] Noman N, Iba H. Enhancing differential evolution performance with local search for high dimensional function optimization [C]//Proceedings of the 2005 conference on genetic and evolutionary computation. Washington DC, America:ACM Press, 2005.

[11] 刘 波,王 凌,金以慧. 差分进化算法研究进展[J]. 控制与决策,2007,22(7):721-729.

[12] 谢晓锋,张文俊,张国瑞,等. 差异演化的实验研究[J]. 控制与决策,2004,19(1):49-52.

[13] 邓泽喜,刘晓冀. 差分进化算法的交叉概率因子递增策略研究[J]. 计算机工程与应用,2008,44(27):33-36.

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

基于差分进化算法的K-Means算法改进

作者：[刘莉莉](#)，[曹宝香](#)，[LIU Li-li](#)，[CAO Bao-xiang](#)

作者单位：[曲阜师范大学 信息科学与工程学院, 山东 日照, 276826](#)

刊名：[计算机技术与发展](#)

英文刊名：[Computer Technology and Development](#)

年，卷(期)：2015(10)

引用本文格式：[刘莉莉](#)，[曹宝香](#)，[LIU Li-li](#)，[CAO Bao-xiang](#) [基于差分进化算法的K-Means算法改进](#)[期刊论文]-[计算机技术与发展](#) 2015(10)