

基于 K-Means 变异算子的混合 PSO 算法聚类研究

杨晓庆, 左为恒, 李昌春

(重庆大学 输配电装备及系统安全与新技术国家重点实验室, 重庆 400030)

摘要: 提出了基于 K-Means 算子的混合粒子群优化算法聚类, 将 K-Means 算法的局部搜索能力与粒子群优化算法的全局寻优搜索能力相结合, 根据群体适应度变化的情况自适应调整权重, 并对种群中性能较差的粒子进行交叉选择, 能充分挖掘群体本身信息, 又能不断引入附加信息. 数据集仿真实验表明, 该算法有效的克服了传统粒子群优化算法过慢收敛和 K-Means 算法陷入局部收敛的问题, 从而得到更好的聚类效果.

关键词: 聚类分析; K-Means 算法; 粒子群优化算法

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2011)07-0057-04

Hybrid PSO Algorithm Clustering Analysis Based on K-Means Mutation Operator

YANG Xiao-qing, ZUO Wei-heng, LI Chang-chun

(State Key Laboratory of Power Transmission Equipment & System Security and New Technology,
Chongqing University, Chongqing 400030, China)

Abstract: This paper presents a hybrid PSO algorithm based on K-Means operator. It combines the locally searching capability of the K-Means algorithm with the global optimization capability of genetic algorithm, and introduces the K-Means operator into the PSO algorithm. It's a hybrid algorithm using symbolic coding, adaptive mutation, and optimal individual retention policies. Simulation results show that the algorithm has effectively overcomes the slow convergence of PSO algorithm and the locality convergence of K-Means algorithm, in order to can get better clustering.

Key words: cluster analysis; K-Means algorithm; PSO algorithm

1 引言

聚类分析方法是指对样品或指标按其某些属性进行分类, 在同一类中的个体具有较高的相似度, 以实现对数据的分类^[1]. 聚类分析作为数据挖掘中的一个重要研究领域而被广泛应用. 聚类分析的算法有很多种, 主要集中在基于距离的聚类分析. 由于粒子群优化算法(PSO)收敛速度快、参数设置少, 已成为当前一种重要的优化工具, 在聚类问题分析研究中具有全局搜索能力, 但在实际应用中也具有局部收敛, 进化后期收敛速度慢, 精度较差等缺点^[2], 针对这一问题, 文中提出了一种基于 K-Means 算法收敛速度快、局部搜索能力强的优点来

改进传统粒子群优化算法的方法. 在数据集上的仿真实验表明, 文中提出的改进算法明显提高了粒子群算法的全局收敛性能, 并能有效避免粒子群优化算法中的早熟收敛问题, 且具有更快的收敛速度, 得到更好的最优解.

2 聚类分析的数学模型

聚类是将样本数据分类到不同的类或者簇这样的一个过程, 所以同一个簇中的对象有很大的相似性, 而不同簇间的对象有很大的相异性. 设 $X = \{x_1, x_2, \dots, x_n\}$ 是给定 R^c 空间中的有限数据集, 将 n 个数据样本按照它们之间的相似度分为 c ($1 < c < n$) 类, 其聚类的中心:

$$C_r = \frac{1}{n_r} \sum_{i=1}^{n_r} X_i^{(r)}$$

聚类的目标函数:

$$\text{value} = \min \sum_{r=1}^c \sum_{j=1}^{n_i} |X_j^{(i)} - C_i|^2$$

式中, $i = 1, 2, \dots, n_r$; n_r 为属于第 r 类的样本个数; $X_i^{(r)}$ 表示样本 X_i 属于第 r 类; $r = 1, 2, \dots, c$; c 为聚类中心数 ($1 < c < n$); n 为样本数。

3 改进算法基本思路及算法流程

3.1 K-Means 算法思想

K-Means 算法的基本思想描述如下^[3]:

- (1) 任意选择 K 个样本作为初始的聚类中心。
- (2) 计算每个样本与 K 个聚类中心的距离, 并将距离聚类中心最近的样本作为该中心所属的类。
- (3) 计算每一类样本属性的均值作为新的聚类中心。
- (4) 重复(2) ~ (3), 直到(3)不再发生变化。

K-Means 算法是聚类分析中一种基本的划分方法。这种算法用每类的平均值来表示该类, 大大降低了算法计算的复杂性, 提高了运算的速度。但传统的 K-Means 算法对初始聚类中心、孤立点都很敏感, 其受初始选定的聚类中心的影响而过早地收敛于局部最优解。因此, 需要一种能克服上述缺点的全局优化算法。

3.2 粒子群优化算法思想(PSO)

在粒子群算法中, 每个优化问题的解都被抽象为没有质量和体积的微粒, 并将其延伸到 N 维空间, 粒子 i 在 N 维空间里的位置表示为一个矢量, 每个粒子的飞行速度也表示为一个矢量。PSO 首先初始化一群随机粒子, 然后粒子们就追随当前的最优粒子在解空间中搜索, 即通过迭代找到最优解。假设 d 维搜索空间中的第 i 个粒子的位置和速度分别为 $X^i = (x_{i1}, x_{i2}, \dots, x_{id})$ 和 $V^i = (v_{i1}, v_{i2}, \dots, v_{id})$, 在每一迭代中, 粒子通过跟踪粒子本身的最优解和整个种群的最优解来更新自己的速度和位置:

$$V_{i,j}(t+1) = w_{i,j}(t) + c_1 r_1 [p_{i,j} - x_{i,j}(t)] + c_2 r_2 [p_{g,j} - x_{i,j}(t)] \quad (1)$$

$$x_{i,j}(t+1) = x_{i,j}(t) + v_{i,j}(t+1) \quad (2)$$

$$j = 1, 2, \dots, d$$

其中, $p_{i,j}$ 是个体极值, $p_{g,j}$ 是全局最优解, w 为惯性因子, c_1 和 c_2 为正的学习因子, r_1 和 r_2 为 $[0, 1]$ 之间均匀分布的随机数。粒子群优化算法依据粒子自身速度进行惯性运动, 对粒子自身行为进行思考的同

时, 参与群体信息共享与相互合作, 从而在粒子群中寻找最佳位置, 上述相互作用与制约决定了算法的寻优性能。

3.3 基于 K-Means 算子的混合 PSO 算法聚类流程

传统的粒子群优化算法由于受其更新方式的限制, 因此群体在演变过程中对周围环境的利用能力变得越来越弱, 在算法优化的后期粒子几乎处于停滞状态^[4]。针对 PSO 算法的这些缺陷, 文中提出一种改进的粒子群优化算法。算法的基本思想是在针对 PSO 算法在聚类问题分析研究中具有良好的全局搜索能力, 利用 K-Means 算法收敛速度快、局部搜索能力强的优点来改进 PSO 算法的思想, 但其在进化后期有收敛慢、很难求得最优解的缺点。把 K-Means 的局部搜索能力与 PSO 算法的全局搜索能力相结合, 在种群进化过程中, 引入 K-Means 变异算子, 同时, 为了避免出现早熟现象和破坏全局的收敛, 引入自适应算法策略, 将遗传算法的自然选择思想引入到粒子群优化算法中。在每次迭代过程中将整个粒子群按适应值进行排序, 保留群体中最好的一半的速度和位置信息, 群体中较差一半的位置和速度进行交叉变异, 同时保留每个个体所记忆的最优历史值。因此, 基于 K-Means 变异算子的混合 PSO 算法聚类既利用了 PSO 算法追逐当前全局最优优点来保证算法的收敛性, 又利用了遗传算法自然选择、优胜劣汰的机理增强了算法搜索的随机性、加大了搜索范围。基于 K-Means 变异算子的混合 PSO 算法聚类的基本步骤如下:

Step1. 通过对聚类模型的分析, 结合 K-Means 算法, 随机初始化种群中各微粒的位置和速度, 染色体结构为 $X = \{x_1, x_2, \dots, x_n\}$, 其为 $1 * n$ 维行向量, x_l 为染色体第 l 位的基因, $x_l \in \{1, 2, \dots, c\}$, $l = 1, 2, \dots, n$;

Step2. 评价每个微粒的适应度, f 的值越小, 说明这种分类方法的误差越小, 聚类效果越好, 如果以 f 值作为适应度函数值, 在进行 PSO 算法操作时, 对于极少数适应度高的个体容易引起算法提前收敛于局部最优解。为了避免这一缺点, 适应度函数的设计如下:

① 对 f 值由小到大排序, 记录其排序号 index 。

② 计算适应度值: $\text{fitness} = a * (1 - a)^{\text{index}-1}$, a 定义为最好个体的选择概率 $a \in (0, 1)$ 。根据参考文献^[5], 经过仿真验证, 文中取 $a = 0.5$ 。并将当前各微粒的位置和适应度存储在各微粒的 pbest 中, 将

所有的 pbest 中适应值最优的个体的位置和适应度储存于 gbest 中。

Step3. 在数据聚类中,传统 PSO 算法当种群中各个个体的目标函数值趋于一致或者趋于局部最优时,就不易产生新的最优个体结构,导致算法容易陷入局部最优。针对这一问题,当种群中的个体目标函数值趋于一致时,采用 K-Means 算法进行个体基因的指定变异,计算每个样本与 K 个聚类中心的距离,并将距离聚类中心最近的样本的基因变异为其相应的类。

通过对图 1 传统 PSO 算法聚类目标函数收敛曲线的分析,文中采用自适应函数更新惯性权重系数,因此,基于自适应 K-Means 算子的动态惯性权重系数更新公式如下:

$$w = \begin{cases} w_{\min} \frac{(w_{\max} - w_{\min}) * (f - f_{\min})}{(f_{\text{avg}} - f_{\min})} & f \leq f_{\text{avg}} \\ \text{K-Means 变异算子} & f > f_{\text{avg}} \end{cases} \quad (3)$$

式中, $w_{\max}$ 、 $w_{\min}$ 分别表示 w 的最大值和最小值, f 表示粒子当前的目标函数值, f_{avg} 和 $f_{\min}$ 分别表示当前所有微粒的平均目标值和最小目标值,当各微粒的目标值趋于一致或趋于局部最优时,使惯性权重增加,而各微粒的目标值比较分散时,使惯性权重减少,同时对于目标函数值优于平均目标值的微粒,其对应的惯性权重因子比较小,从而保护了该微粒。

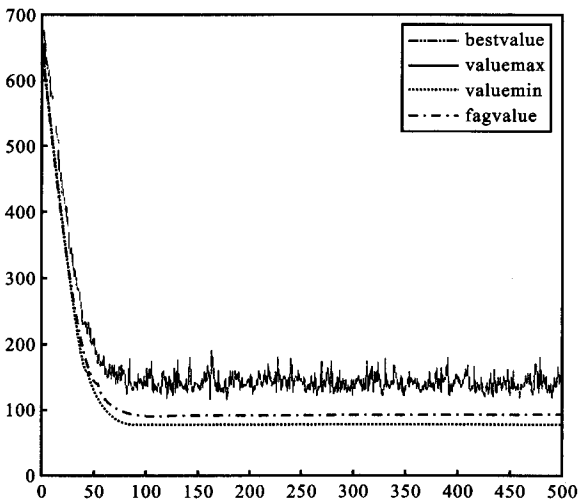


图 1 传统 PSO 算法聚类目标函数收敛曲线

Step4. 用下式更新粒子的速度和位置:

$$\begin{aligned} V_{i,j}(t+1) &= wv_{i,j}(t) + c_1r_1[p_{i,j} - x_{i,j}(t)] + \\ c_2r_2[p_{g,j} - x_{i,j}(t)]x_{i,j}(t+1) &= x_{i,j}(t) + v_{i,j}(t+1) \\ j &= 1, 2, \dots, d \end{aligned} \quad (4)$$

Step5. 重新评价每个微粒的适应度,将其适应度值与经历过的最好位置作比较,如果较好,则将其作为当前的最好位置,比较当前所有的 pbest 和 gbest 的值,更新 gbest 的值,并按适应值的大小对种群排序,保留适应值较好的一半,将适应值较差的一半按下式选择较差。

$$\text{child}(x) = p \cdot \text{parent}_1(x) + (1 - p) \cdot \text{parent}_2(x) \quad (5)$$

Step6. 若算法满足停止条件(通常为预设的运算精度或迭代次数),搜索停止,输出结果。否则返回 Step2 继续搜索。

4 仿真实验

为了检验算法的有效性,对 K-Means 算法、传统粒子群优化算法、文中算法进行对比实验。在 Matlab 实验平台上分别编写算法程序,导入数据进行试验。实验数据有 3 组:第 1 组实验数据 Data1 包含 12 个样本,分为 4 类,每个样本有 2 个属性;第 2 组实验数据,数据集是 Iris,其中包含 150 个样本,分为 3 类,每类 50 个数据,每个样本包含 4 个属性;第 3 组实验数据 Data3 包含 2400,分为 7 类,每个样本包含 2 个属性。其中选取粒子数目 N 为 40,最大权重 $w_{\max}$ 为 0.9,最小权重 $w_{\min}$ 为 0.6。对 3 种算法、3 组数据进行 10 次运算,结果比较如表 1 所示。

根据表 1 的实验数据结果可见,K-Means 算法由于对初始聚类中心选取的敏感,并不是每次都能得到最优解,但其对数据的聚类运算速度快;传统 PSO 算法对小规模数据聚类可以达到最优解,但其收敛速度较慢,对于稍大规模的数据,其很难得到最优解,特别在数据 3 里面传统 PSO 算法没有一次得到最优解;文中算法对三组数据的聚类都能得到最优解,而且其收敛速度也较快,有效克服了传统 PSO 算法收敛慢的缺点和 K-Means 算法对初始聚类中心的敏感性。

5 结束语

文中结合 PSO 算法的全局搜索寻优能力和 K-Means 算法局部搜索能力,提出了一种基于 K-Means 算法收敛速度快、局部搜索能力强的优点来改进传统粒子群优化算法的方法。仿真实验表明,该算法能克服 PSO 算法收敛慢、K-Means 算法对初始聚类中心敏感陷入局部收敛的缺点,可以很好的对数据进行聚类分析,得到全局最优解,为聚类分析研究提供了一个新的思路。

表 1 K-Means 算法、传统 PSO 算法、文中算法聚类结果比较

数据集	算法	达到最优解 平均迭代次数	达到最优解 平均时间	最优解 最大值	最优解 均值	最优解 最小值	最优 解值	达到最优 解次数
K-Means 算法	数据 1	2.6	0.000693	12.833	9.700	5.333	5.333	4
	数据 2	7.1	0.018835	145.279	98.478	78.94	78.94	7
	数据 3	10.8	0.511108	20114.1	15698.2	11594.1	11594.1	5
传统 PSO 算法	数据 1	24.1	0.013357	12.833	6.317	5.333	5.333	8
	数据 2	161.3	0.978232	78.94	78.94	78.94	78.94	10
	数据 3	455.1	23.23743	206181	205238.6	204038	204038	0
文中算法	数据 1	3	0.004770	5.333	5.333	5.333	5.333	10
	数据 2	8.3	0.239138	78.94	78.94	78.94	78.94	10
	数据 3	7.5	1.711048	11594.1	11594.1	11594.1	11594.1	10

参考文献:

- [1] 元昌安,邓松,李文敬,等.数据挖掘原理与 SPSS Clementine 应用宝典[M].北京:电子工业出版社,2009:268-310.
- [2] 高翔,唐普英.设立禁区的多粒子群优化算法[J].计算机工程与应用,2010,46(10):39-40.
- [3] 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.
- [4] 童亚拉.基于自适应混沌粒子群的 Web 搜索结果聚类研究[J].微电子学与计算机,2010,1(27):174-176.
- [5] 杨淑莹.图像模式识别—VC++ 技术实现[M].北京:清华大学出版社,2005:241-266.
- [6] 龚纯,王正林.精通 MATLAB 最优化计算[M].北京:电子工业出版社,2009:313-318.

作者简介:

杨晓庆 女,(1985—),硕士研究生.研究方向为智能控制、控制理论与控制工程.

左为恒 男,(1961—),副教授.研究方向为智能控制、控制理论与控制工程.

李昌春 女,(1965—),副教授.研究方向为智能控制、控制理论与控制工程.

(上接第 56 页)

- [2] 国蓉,何镇安,王伟.被动声探测技术与弹着点定位方法综述[J].电声技术,2010,34(11):48-52.
- [3] Howell B P, Wood S, Koksall S. Passive Sonar Recognition and Analysis Using Hybrid Neural Networks. OCEAN'2003 [EB/OL]. [2010-03-13]. <http://my.fit.edu/~swood/PassiveSonar.pdf>
- [4] 张宝军.基于改进 RBF 神经网络的 TDOA/AOA 定位算法[J].微电子学与计算机,2009,26(7):183-186.
- [5] 梁艳,靳东明.基于 RBF 神经网络参数优化的亲属优先遗传算法[J].微电子学与计算机,2009,26(7):17-20.
- [6] 王成,李少洪,王鑫全,等.测时差被动定位算法的研究[J].系统工程与电子技术,2001,23(11):9-12.
- [7] 张德丰. MATLAB 神经网络应用设计[M].北京:机械工业出版社,2009:157-160.

作者简介:

国蓉 女,(1973—),博士,副教授.研究方向为智能检测、微机电引信、数字信号处理.

何镇安 男,(1985—),硕士研究生.研究方向为人工智能.