改进的粒子群优化模糊 C 均值聚类算法

温重伟, 李荣钧
(华南理工大学 工商管理学院, 广州 510640)

摘 要: 针对传统模糊 C 均值聚类算法(FCM)存在对初值敏感和易陷入局部收敛的缺陷,利用改进的粒子群算法对 FCM 进行优化,提出一种新的模糊 C 均值聚类算法 Improved PSOFCM,并建立基于熵的聚类有效性函数,对聚类算法的性能进行客观评价。数据集实验表明,Improved PSOFCM 算法不仅能克服传统 FCM 算法的不足,而且在聚类正确率和有效性上也优于基于粒子群与基于遗传优化的 FCM 算法。

关键词: 模糊 C 均值聚类; 粒子群优化; 熵; 聚类有效性
中图分类号: O159; TP18 **文献标志码**: A **文章编号**: 1001-3695(2010)07-2520-03
doi:10.3969/j.issn.1001-3695.2010.07.033

Fuzzy C-means clustering algorithm based on improved PSO

WEN Zhong-wei, LI Rong-jun
(*College of Business Administration, South China University of Technology, Guangzhou* 510640, *China*)

Abstract: Traditional FCM clustering algorithm includes the problems of local optimal and sensitivity to initial values. The improved PSO algorithm was used to optimize FCM. This paper proposed a new fuzzy C-means clustering algorithm Improved PSOFCM and constituted a clustering efficiency function based on Shannon entropy to evaluate the performance of clustering algorithm in the impersonal way. Numerical examples illustrate that the Improved PSOFCM can overcome the deficiency of FCM, and have better clustering accuracy and efficiency than FCM based on PSO and GA.

Key words: fuzzy C-means(FCM) clustering; particle swarm optimization(PSO); entropy; clustering efficiency

聚类就是在没有先验知识的情况下,仅靠事物间的相似性作为类属划分的准则对数据进行分类,使得同一类中的数据相似性尽量高,不同类中的数据相似性尽量低。传统的聚类分析要求把数据集中的每一点都精确地划分到某个类中,即所谓的硬划分。但现实中的大多数事物往往具有模糊性,即事物间没有明确的界限,不具有非此即彼的性质,所以模糊聚类的概念更适合事物的本质,能更客观地反映现实。模糊聚类方法认为每一个样本与各个聚类中心都有一个隶属关系。表达了样本类属的模糊性,能更客观地反映现实世界,从而成为聚类分析研究的主流。目前,模糊 C 均值(FCM)聚类算法是应用最广泛的一种模糊聚类算法。但是,传统的 FCM 算法存在两个致命的缺陷:a)算法的性能依赖于初始聚类中心的选取,同时聚类的效果受初始值的影响较大;b)FCM 算法在迭代寻找最优解的过程中使用的是梯度下降的方法,导致不可避免地会陷于局部最优值。因此,针对算法对初值敏感和易陷入局部收敛,近年来,结合遗传算法(genetic algorithm, GA)的模糊聚类分析成为了研究的新方向,文献[1]给出了一种基于遗传算法的模糊聚类方法 GGA,改进了模糊聚类的质量。文献[2]先运用遗传算法得到聚类中心,然后用改进的 FCM 算法求出最优解,在通信信号的星座聚类中得到了很好的效果。文献[3]提出了一种基于改进的遗传算法的聚类方法 Improved GAFCM,在标准遗传算法的基础上,引入了避免近亲繁殖的交叉方法和大变异操作,避免陷入局部最小值,将该算法运用于 Iris 数据的聚类,取得了较好的聚类效果。

作为另一类智能计算方法的代表,与进化算法相比,粒子群算法(PSO)由于算法概念简单、实现容易,短短几年时间便获得了很大发展。相对 GA 而言,PSO 不需要进行交叉和变异,操作较为简单,而且在迭代过程中,粒子运动的思路与人类决策相似,易于理解。另外,在 GA 中,染色体互相共享信息,所以整个种群的移动是比较均匀地向最优区域移动,在 PSO 中,只有最优粒子释放信息给其他粒子,整个搜索更新过程跟随当前最优解的过程,属于单向的信息流动。因此,与 GA 比较,在大多数的情况下,所有的粒子可能更快地收敛于最优解。同时,研究表明,在 PSO 算法最具潜力的应用领域中,分类和模式识别就在其中[4]。因此,结合粒子群算法对模糊 C 均值聚类的迭代过程进行优化,是克服 FCM 算法缺陷的一个更有效的方法[5],文献[5]对此作了研究。但是,使用传统 PSO 算法易陷入局部极值而出现早熟收敛,为了对其进行改进,本文引入了一种带有邻域操作的 PSO 模型,将邻域极值也作为粒子进化的一个信息来源,有效地避免了传统 PSO 算法可能出现的早熟收敛现象;利用这种改进粒子群算法对 FCM 进行优化,提出一种新的基于改进粒子群优化的模糊 C 均值聚类算法,并通过基于熵的聚类有效性函数对算法性能进行测试。

1 模糊 C 均值聚类算法

FCM 是 Bezdek 于 1981 年提出的,是目前广泛采用的一种聚类算法。算法用隶属度矩阵给出了每个样本隶属于某个聚类的程度,即使对于很难明显分类的变量,模糊 C 均值聚类也

收稿日期: 2009-12-26; **修回日期**: 2010-01-30

作者简介:温重伟(1986-),男,广东韶关人,硕士研究生,主要研究方向为管理科学与工程、数据挖掘技术、金融与财务决策(james_wen2004@qq.com);李荣钧(1946-),男,教授,博导,博士,主要研究方向为运筹学、优化理论与模糊技术、投资决策与风险管理.

能得到较为满意的效果[6]。考虑一个具有 n 个元素的数据集 $X=\{x_1,x_2,\cdots,x_n\}$,每个元素包含 d 个属性。聚类数为 $c(2\leqslant c\leqslant n)$,$W=\{w_1,w_2,\cdots,w_c\}$ 为聚类中心。模糊聚类的每个元素都不能严格地被划分入某一类中,因此令 u_{ik} 表示第 k 个元素属于第 i 类的隶属度,其中,$\sum_{i=1}^{c}u_{ik}=1$,$u_{ik}\in[0,1]$。

模糊 C 均值聚类的目标函数定义如下:

$$\min J_m(U,W)=\sum_{k=1}^{n}\sum_{i=1}^{c}u_{ik}^m d_{ik}^2 \tag{1}$$

其中:$d_{ik}=\|x_k-v_i\|$ 表示元素 x_k 与类中心 w_i 之间的欧式距离,$m\geqslant 1$ 为影响隶属度矩阵模糊化程度的指数权重,通常取 $m=2$。模糊 C 均值聚类算法的思想就是迭代调整 (U,W),使得目标函数最小。迭代的步骤如下:

a)取定聚类数目 c 和权重 m,随机生成聚类中心矩阵 $W^{(0)}$,并令迭代次数 $l=0$。

b)计算隶属度矩阵 U 为

$$u_{ik}^{(l)}=\begin{cases}1/\sum_{j=1}^{c}(d_{ik}/d_{jk})^{\frac{2}{m-1}} & d_{ik}>0\\ 1 & d_{ik}=0\end{cases} \tag{2}$$

c)修正聚类中心 W 为

$$w_i^{(l+1)}=\sum_{k=1}^{n}(u_{ik}^{(l)})^m x_k/\sum_{k=1}^{n}(u_{ik}^{(l)})^m \tag{3}$$

d)对于给定阈值 $\varepsilon>0$,若目标函数 $\|J_m^{(l)}-J_m^{(l-1)}\|\leqslant\varepsilon$,则算法终止;否则 $l=l+1$,转到 b)。

FCM 算法迭代终止后,模糊隶属度矩阵 $U^{(l)}$ 对应样本 X 的模糊划分,可以采用最大隶属原则使聚类清晰化,即对于 $U^{(l)}$ 的第 k 列,若 $u_{i_0k}=\max_{1\leqslant i\leqslant c}(u_{ik})$,则 x_k 属于第 i_0 类。FCM 的最优聚类中心为 $W^{(l)}=\{w_1^{(l)},w_2^{(l)},\cdots,w_c^{(l)}\}$,最优值为 $J_m^{(l)}$。

2　基于粒子群算法的模糊 C 均值聚类

2.1　粒子群优化算法(PSO)

PSO 算法是 Kennedy 和 Eberhart 受到鸟群觅食行为的启发,于 1995 年提出并用于解决复杂优化问题[7]。算法采用速度—位移搜索模型,每个粒子 s_i 代表解空间的一个候选解,粒子具有两个属性,即决定优劣程度的适应值 f_i(fitness value)与决定粒子移动方向和距离的速度 v_i,其中,适应值根据实际问题所设计的适应值函数来计算。PSO 首先初始化一群随机粒子,通过迭代找到最优解。每一次迭代,粒子通过动态跟踪两个极值来更新其速度和位置:第一个是粒子本身到当前迭代次数为止搜索产生的最优解,称为个体极值 p_i;第二个是整个种群目前的最优解,记为全局极值 g。在这两个极值的指引下,粒子根据某种策略来更新其速度和位置,直到满足终止条件,搜索得到最优解。

PSO 算法是基于群体智能理论的优化算法,通过群体中粒子间的合作与竞争产生的群体智能指导优化搜索。正如引言中提到的,与进化算法比较,PSO 保留了基于种群的全局搜索策略,其采用的速度—位移模型操作简单,避免了复杂的遗传操作。它特有的记忆使其可以动态跟踪当前的搜索情况,调整其搜索策略,与进化算法比较,粒子群优化算法是一种更高效的并行搜索算法[8]。而且包括 Angeline、Shi 和 Eberhart 在内的学者经过大量的实验研究发现,在解决一些函数优化问题时,PSO 算法在优化速度和精度上均比遗传算法有一定的改善[9,10]。

2.2　基于粒子群算法的模糊 C 均值聚类

如前所述,FCM 算法由于使用梯度下降方法寻找最优解,存在对初始值敏感和容易陷入局部最优解的缺陷。利用 PSO 的优化搜索能力对 FCM 算法进行改进,无论初始值如何选取,都能保证得到问题的全局最优解。在使用 PSO 算法进行优化求解时,关键是对编码、适应值函数、粒子速度—位移更新策略三个部分的设计。下面就算法设计中,对这些主要问题的解决进行详细说明:

a)个体确定与粒子编码。聚类算法的关键是确定聚类中心,因此可以选取聚类中心作为种群中的个体。设样本维数为 d,聚类中心数为 c,则每个粒子实际上是一个 $d\times c$ 维的实向量。粒子采用实数编码的方式,编码长度为 $d\times c$,具体结构如图 1 所示。

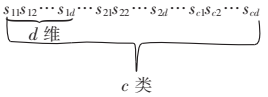

图 1　粒子编码结构

b)适应值函数。对于 FCM 算法而言,最优聚类结果即目标函数式(1)取得最小值时对应的结果。由于粒子群算法一般取适应值的最大值对应最优解,可以把 FCM 目标函数的倒数定义为适应值函数:

$$f=1/(J_m+\varepsilon) \tag{4}$$

其中:ε 为充分小的正实数,实际计算中一般取 $\varepsilon=10^{-10}$。

c)速度—位移更新机制。粒子的速度—位移更新机制决定了算法的收敛速度和精度,Eberhart 和 Kennedy 最初提出的更新策略如下[11]:

$$v_i=v_i+c_1r_1(p_i-s_i)+c_2r_2(g-s_i) \tag{5}$$

$$s_i=s_i+v_i \tag{6}$$

其中:c_1、c_2 为学习因子,一般取 $c_1=c_2=2$;r_1、r_2 是均匀分布在(0,1)区间的随机数。粒子通过上述更新策略不断跟踪个体极值和全局极值进行搜索,直到满足终止条件为止。在实际优化问题中,为了缩短算法的收敛时间并保证精度,往往希望先使用全局搜索把最优解空间划定在某个小的区域,然后在该区域采用局部搜索以获得高精度的解。因此,Shi 与 Eberhart 在式(5)中引入惯性权重 ω[12]:

$$v_i=\omega v_i+c_1r_1(p_i-s_i)+c_2r_2(g-s_i) \tag{7}$$

ω 较大则算法具有较强的全局搜索能力;反之,则算法倾向于进行局部搜索。通常将 ω 由初始的最大值 $\omega_{\max}=0.9$ 随着迭代次数的增加线性递减至最小值 $\omega_{\min}=0.4$,即

$$\omega=\omega_{\max}-\text{iter}\times(\omega_{\max}-\omega_{\min})/\text{iter}_{\text{total}} \tag{8}$$

其中:iter 为当前迭代次数,$\text{iter}_{\text{total}}$ 为累计迭代次数。

实际上,这种速度—位移更新策略仍然存在问题,若迭代过程出现这样的情况:个体最优位置或全局最优位置离得较近且恰好为局部最优解时,所有粒子都朝此方向进化,产生新的较好个体的可能性较小,从而使得该算法易出现早熟收敛。为了避免这种情况,本文借鉴 Suganthan 的带有邻域操作的 PSO 模型[13],定义粒子 s_i 的邻域极值为 l_i,把该极值也作为粒子进化的一个信息来源。在优化的初始阶段,将邻域定义为每个粒子自身,随着迭代次数的增加,将邻域范围逐步扩展到包含所有粒子,这样就避免了上述情况出现而导致早熟收敛,新的更新策略调整为

$$v_i=\omega v_i+c_1r_1(p_i-s_i)+c_2r_2(g-s_i)+c_3r_3(l_i-s_i) \tag{9}$$

d)终止条件。以下两种情况只要满足其一,算法终止,即最优解对应的目标值保持不变或变动小于阈值 E_{rr} 的持续迭代次数达到设定值 $iter_{stable}$,或者迭代次数已达到设定的最大次数 $iter_{max}$。

综上分析,把基于粒子群算法的模糊 C 均值聚类记为 Improved PSOFCM,具体步骤如下:

a)对算法参数赋值,包括聚类数目 c、粒子种群规模 $swarm_{size}$、允许的最大速度 v_{max}、最大迭代次数 $iter_{max}$、最优解改变量阈值 E_{rr} 及其迭代次数阈值 $iter_{stable}$;

b)在样本各属性值的范围内,按照编码原则随机生成初始种群,每个粒子代表各类的聚类中心;

c)根据式(4)计算初始种群中个体的适应值;

d)根据式(9)计算粒子的速度,并通过式(6)更新粒子的位移,迭代次数加 1;

e)计算种群中的个体适应值,若满足终止条件,则算法结束,否则,转 d)继续进行。

2.3 基于熵的聚类有效性分析

聚类有效性分析是评价聚类效果的重要手段。文献[14]给出了一种基于香农信息熵的聚类有效性函数,对于给定的聚类中心数 c 和模糊隶属度矩阵 U,基于熵的聚类有效性函数 $HP(U;c)$ 定义为[14]

$$HP(U;c)=H_1(U;c)-H_2(U;c) \tag{10}$$

$$H_1(U;c)=-\sum_{k=1}^{n}\sum_{i=1}^{c}\frac{u_{ik}}{n}\ln\frac{u_{ik}}{n} \tag{11}$$

$$H_2(U;c)=-\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}/\left(c\sum_{i=1}^{n}u_{ik}\right)\ln\left(u_{ik}/\left(c\sum_{k=1}^{n}u_{ik}\right)\right) \tag{12}$$

评价准则为:$HP(U;c)$ 的函数值越小,则聚类效果越好。

3 实验与结果分析

为了测试算法的性能,选择 UCI 机器学习数据库中的 Iris 和 Wine 数据集进行实验,并与传统 FCM 算法、文献[3]中提出的改进型遗传算法的 FCM 算法 Improved GAFCM 以及文献[5]的粒子群优化 FCM 算法 PSOFCM 作比较。其中,Iris 数据集由 150 个四维向量样本组成,共分为三个种类,每一个种类有 50 个样本;Wine 数据集由 178 个 13 维向量样本组成,也分为三个种类,各类样本数目为 59、71 和 48,这两个数据集常被用来检验聚类算法的性能。设置粒子种群规模 $swarm_{size}=50$,聚类数目 $c=3$,允许的最大速度 $v_{max}=2$,最大迭代次数 $iter_{max}=100$,最优解改变量阈值 $E_{rr}=0.01$ 以及迭代次数阈值 $iter_{stable}=5$。分别对各算法运行 10 次,取各指标的平均值,结果如表 1 所示。

表 1 模糊 C 均值聚类算法的性能比较

数据集	算法	$J_m(U,W)$ 均值	$J_m(U,W)$ 方差	正确率/%	迭代数	$HP(U;c)$
Iris	FCM	80	8.46	88.7	21	0.004
	文献[3]	64	0.48	90.0	180	-0.002
	文献[5]	60	0.46	92.0	100	-0.003
	本文	44	0.43	93.3	100	-0.004
Wine	FCM	618	70.02	69.1	46	0.768
	文献[3]	589	25.15	75.3	200	0.712
	文献[5]	585	17.36	76.4	100	0.658
	本文	563	12.25	79.8	100	0.625

由表 1 可以看出,基于粒子群算法的模糊 C 均值聚类 Improved PSOFCM 在所有指标上都优于其他算法。其中,传统 FCM 算法由于采用了梯度算子,函数值下降非常迅速,但马上又陷入了局部最小值,这是使用梯度算子很难避免的结果;而且最优值的方差较大,说明最优解不稳定,对初值选取较敏感,通过聚类有效性函数的计算结果也说明,传统 FCM 算法的聚类效果相对较差。与采用遗传优化和粒子群优化的 FCM 算法相比,由于算法都采用了智能优化技术进行处理,聚类正确率较高,且最优解方差较小,说明算法的全局搜索能力较好;相比之下,PSO 算法结构简单,与基于遗传算法的 FCM 相比,迭代次数大为减少,且改进的 PSO 算法较传统 PSO 算法具有更强的全局搜索能力,聚类有效性函数值也表明,Improved PSOFCM 的聚类效果最优。

4 结束语

综上所述,FCM 算法存在对初值敏感、易陷入局部极值的缺陷,结合智能优化算法的 FCM 能够有效地解决这个问题。考虑到传统 PSO 算法有可能出现早熟收敛的情况,本文引入邻域极值作为粒子进化的一个信息来源,提出一种基于改进粒子群优化的 FCM 算法 Improved PSOFCM。实验结果表明,Improved PSOFCM 算法不仅弥补了 FCM 算法的缺陷,而且在聚类效果和搜索性能上也优于其他两种基于智能优化的 FCM 算法。

参考文献:

[1] HALL L O, OZYURT B, BEZDEK J C. Clustering with a genetically optimized approach[J]. IEEE Trans on Evolutionary Computation, 1999, 3(2): 103-112.

[2] 宋娇, 葛临东. 一种遗传模糊聚类算法及其应用[J]. 计算机应用, 2008, 28(5): 1197-1199.

[3] 殷晓明, 顾幸生. 一种基于改进型遗传算法的模糊聚类[J]. 华东理工大学学报: 自然科学版, 2006, 32(7): 849-851.

[4] 胡一波. 求解约束优化问题的几种智能算法[D]. 西安: 西安电子科技大学, 2009.

[5] 蒲蓬勃, 王鸽, 刘太安. 基于粒子群优化的模糊 C-均值聚类改进算法[J]. 计算机工程与设计, 2008, 29(16): 4277-4279.

[6] TANG L, HUANG P Z, XIE W X. A new method of FCM considering the distribution of data[J]. Geomatic and Information Science of Wuhan University, 2003, 28(4): 476-479.

[7] KENNEDY J, EBERHART R C. Particle swarm optimization[C]// Proc of IEEE International Conference on Neural Networks. 1995: 1942-1948.

[8] 章万国, 周驰, 高海兵, 等. 粒子群优化算法[J]. 计算机应用研究, 2003, 20(12): 7-11.

[9] ANGELINE P J. Evolutionary optimization versus particle swarm optimization: philosophy and performance difference[C]//Proc of the 7th Annual Conference Center on Evolutionary Programming. London, UK: Springer-Verlag, 1998: 601-610.

[10] SHI Y H, EBERHART R C. Experimental study of particle swarm optimization[C]//Proc of SCI Conference. 2000.

[11] EBERHART R C, KENNEDY J. A new optimizer using particle swarm theory[C]//Proc of the 6th International Symposium on Micro Machine and Human Science. Nagoya: IEEE Press, 1995: 39-43.

[12] SHI Y H, EBERHART R C. A modified particle swarm optimizer [C]//Proc of IEEE International Conference on Evolutionary Computation. Piscataway, NJ: IEEE Press, 1998: 69-73.

[13] SUGANTHAN P N. Particle swarm optimizer with neighborhood operator[C]//Proc of Congress on Evolutionary Computation. Piscataway, NJ: IEEE Press, 1999: 1958-1962.

[14] 雷鸣. 模糊聚类新算法的研究[D]. 天津: 天津大学, 2006.