基于粒子群优化的模糊 C 均值聚类算法*

王宇钢
(辽宁工业大学 机械工程与自动化学院,辽宁 锦州 121000)

摘　要: 针对模糊 C 均值聚类算法(FCM)存在对初始聚类中心敏感,易陷入局部最优解的不足,将改进的粒子群聚类算法与 FCM 算法相结合,提出了一种基于粒子群优化的模糊 C 均值聚类算法。该算法对粒子群初始化空间及粒子移动最大速度进行优化,同时引入环形拓扑结构邻域,提高粒子群聚类算法的全局搜索能力。对 UCI 中 3 个数据集进行仿真实验,结果表明提出的基于粒子群优化的模糊 C 均值聚类算法相比 FCM 算法和基本粒子群聚类算法具有更好的聚类效率和准确性。

关键词: 聚类; 粒子群优化; 模糊 C 均值聚类算法; 粒子群聚类算法
中图分类号: TP301　　**文献标识码:** A　　DOI: 10.19358/j.issn.2096-5133.2018.08.009
引用格式: 王宇钢. 基于粒子群优化的模糊 C 均值聚类算法[J]. 信息技术与网络安全,2018,37(8):36-39,44.

Fuzzy C-means clustering algorithm based on particle swarm optimization

Wang Yugang
(School of Mechanical Engineering and Automation, Liaoning University of Technology, Jinzhou 121000, China)

Abstract: FCM algorithm is sensitive to initial clustering center and liable to be trapped in a local optimum solution. Combining with the improved PSO algorithm, a fuzzy C-means clustering algorithm based on particle swarm optimization was proposed. The algorithm optimizes the particle swarm initialization space and the maximum velocity of particle, and adopts the ring topology neighborhood. The method improved the global search capability of particle swarm clustering algorithm. The experiment results of UCI data set demonstrate that the proposed algorithm has better clustering validity and accuracy than FCM and particle swarm clustering algorithm.

Key words: clustering; particle swarm optimization; fuzzy C-means clustering algorithm; particle swarm clustering algorithm

0　引言

随着大数据、云计算等技术的迅猛发展,聚类分析已成为数据挖掘的主要研究手段之一。为符合人类的认知,研究员将模糊集理论引入聚类分析中,提出了模糊 C 均值聚类算法(Fuzzy C-means Clustering Algorithm,FCM)。经典 FCM 算法由于是一种局部最优搜索算法,存在对初始聚类中心敏感、易于陷入局部最优解的缺陷,限制了算法的应用[1-2]。因此,学者尝试通过各种智能算法对经典 FCM 算法进行改进。粒子群优化算法(Particle Swarm Optimization, PSO)作为群体智能算法的代表,依靠个体之间的简单交互作用在群体内自组织搜索,具有很强的学习能力和适应性[3]。一些学者利用 PSO 算法克服传统 FCM 算法的缺陷,将 PSO 算法与 FCM 算法融合已成为近年来的研究热点[4]。

文献[5]针对 FCM 算法用于高维数据样本聚类时效果较差的不足,提出一种基于粒子群的 FCM 聚类算法。该算法在满足 FCM 算法对隶属度限制条件的前提下,根据样本与聚类中心间距离重新分布了隶属度,并通过比较样本与各聚类中心距离加速最优粒子收敛。文献[6]对初始聚类中心和模糊加权指数进行粒子编码,通过粒子群优化算法搜索最优的适应度值及模糊加权指数,经人工数据集与 UCI 数据集实验,证明该方法比传统的 FCM 算法和粒子群聚类算法的聚类准确性和稳定性都有提高。文献[7]将基于直觉模糊的粒子群算法(IFPSO)和 FCM 算法混合,利用犹豫度属性参数寻找目标函数与聚类中心的相似性,对高维数据集进行聚类分析取得较好效果。文献[8]提出一种基于惯性指数权重的粒子群聚类算法(ACL-PSO)。将改进的 PSO 算法与 FCM 算法相结合,改善 FCM 算法易于陷入局部最优解的缺陷,对 UCI 数据库中标准数据集进行测试,结果显示了该算法的有效性。

为克服 FCM 算法缺陷,提高聚类质量,本文对基本粒子群聚类算法进行改进,并与 FCM 算法结合,提出了一种改进的粒子群优化模糊 C 均值聚类算法(Improved Fuzzy C-mean Clustering Algorithm Based on Parti-

* 基金项目:辽宁省自然科学基金资助项目(20170540445)

cle Swarm Optimization，IFCM-PSO)。首先通过选择合理的粒子初始化空间，降低对初始聚类中心的敏感度，提高收敛速度；其次通过优化参数粒子运动最大速度以及引入环形拓扑结构的邻域，解决粒子群聚类算法易早熟收敛的缺陷。选取 UCI 数据库中 3 个真实数据集 IRIS、WINE 和 Breast Cancer Wisconsin (BCW) 进行仿真实验，以验证该算法的有效性。

1 模糊 C 均值聚类算法(FCM)

分为 L 个类簇的数据样本集合 $X=\{x_1,x_2,\cdots,x_n\}\in \mathbf{R}^p$，$n$ 为样本个数，p 为样本空间维数，L 介于 $2\sim n$ 之间。FCM 算法采用误差平方和函数作为目标函数，其定义式为：

$$\min J(U,V)=\sum_{i=1}^{L}\sum_{j=1}^{n}u_{ij}^{m}d_{ij}^{2} \tag{1}$$

$$u_{ij}=1/\sum_{k=1}^{L}\left(\frac{d_{ij}}{d_{kj}}\right)^{2/(m-1)} \tag{2}$$

$$v_i=\sum_{j=1}^{n}u_{ij}^{m}x_j/\sum_{j=1}^{n}u_{ij}^{m} \tag{3}$$

其中，$d_{ij}=\|x_j-v_i\|$ 为样本与聚类中心间距离，通常为欧式距离；m 为模糊加权指数；u_{ij} 表示数据集 X 中的第 j 个样本对第 i 类的隶属程度($0<u_{ij}<1$)；v_i 表示各个聚类中心。

隶属度 u_{ij} 应满足约束条件：

$$\sum_{i=1}^{L}u_{ij}=1,j=1,2,\cdots,n \tag{4}$$

FCM 算法是以误差平方和为准则函数的一种逐点迭代聚类算法。通过式(2)和式(3)迭代计算隶属度矩阵 U 和聚类中心 V，使目标函数 $J(U,V)$ 的取值不断减小。当准则函数会聚时，获得数据样本的最终聚类结果，即模糊划分后的隶属度矩阵 U 和聚类中心 V。

2 基本粒子群聚类算法

2.1 粒子群优化算法

在粒子群优化算法中，每个粒子 s_i 抽象为一个个体，种群就是由这些粒子构成的，所求问题的解就是粒子在空间中的最优位置。在每次迭代计算过程中，根据所有粒子的适应值评价每个粒子的极值当前最优位置 p_i 和群体全局最优位置 g。依靠两个位置极值，粒子更新其移动速度和位置，直至收敛到空间位置的最优解。

目前普遍采用的粒子速度和位移更新形式为：

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

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

其中，c_1、c_2 为学习因子，一般取 $c_1=c_2$；r_1、r_2 是 $[0,1]$ 之间的随机数；w 为惯性权重，取值限定在 $[w_{\min},w_{\max}]$ 之间。在迭代过程中，惯性权重通常采用线性递减方式由最大值变为最小值，即：

$$w=w_{\max}-\text{iter}\times(w_{\max}-w_{\min})/\text{iter}_{\text{totle}} \tag{7}$$

其中，iter 为当前迭代次数，$\text{iter}_{\text{totle}}$ 为最大迭代次数。

2.2 FCM-PSO 算法

为了实现传统聚类方法缺陷的突破，研究人员尝试将粒子群优化算法与传统聚类算法相结合，通过 PSO 算法的全局寻优能力和分布式随机搜索特性解决传统聚类算法易陷入局部最优和对初值敏感的问题。将聚类作为一种优化问题实现对数据集的近似最优划分。基本粒子群聚类算法的流程如下：

(1) 给定聚类的数目，初始化聚类中心矩阵，并赋值给各个粒子，随机产生粒子的初始速度。

(2) 对每个粒子计算隶属度，更新所有的聚类中心，计算各个粒子的适应值，更新个体极值。

(3) 根据各个粒于的个体极值，找出全局极值和全局极值位置。

(4) 根据粒子群优化算法的速度公式更新粒子的速度，并把它限制在最大速度内。

(5) 根据粒子群优化算法的位置公式更新粒子的位置。

(6) 若不满足终止条件，返回步骤(2)继续迭代计算；若满足终止条件，则输出最优粒子的位置即最优分类中心矩阵。

目前，将 FCM 算法与 PSO 算法相融合的聚类算法(Fuzzy C-Mean Clustering Algorithm Based on Particle Swarm Optimization，FCM-PSO)已成为基本粒子群聚类算法的一种主要研究形式[9]。该方法将每个粒子表示为一种聚类中心的选取方式，应用 FCM 算法的目标函数计算各粒子的适应值，作为对应聚类中心聚类效果的评判依据，算法收敛后输出粒子的全局最优位置，即最优聚类中心。

3 改进的粒子群优化模糊 C 均值聚类算法

3.1 粒子群聚类算法的改进

(1) PSO 算法通常将粒子初始值均匀分布于 $[0,1]$ 之间，而非在粒子的最优解的附近空间，这将使粒子搜寻最优解的迭代时间增加，聚类的效果变差[10]。本文将样本聚类中心作为种群个体，因此粒子的最优解空间即为样本的分布空间。将粒子的初始位置随机分布于取值范围 $[X_{\min},X_{\max}]$，$X_{\min}$、$X_{\max}$ 分别为样本每维最小值和最大值组成的向量。这样初始化的粒子在接近最优解的搜索空间开始进化运算，可有效缩短收敛时间，提高聚类质量。

(2) 最大速度 v_{max} 决定粒子在一次迭代计算中的最大移动距离，v_{max} 过大则易使粒子错过最优解，过小则会使粒子易陷入局部最优解。因此，通常将粒子最大速度设为一个常数。然而，在样本各维取值存在较大量纲差异时，由于各维空间取值范围不同，将粒子的 v_{max} 在样本各维空间均设定为一个常数，显然易出现错过最优解或陷入局部最优解的情况，结果影响算法的全局收敛性。本文对粒子在样本空间每一维都定义一个最大速度，最大速度 v_{max} 根据样本每维变化的取值范围设定。

$$v_{max} = \lambda(X_{max} - X_{min}) \tag{8}$$

其中，λ 为常数。

(3) 在实际应用中，PSO 算法仍易出现早期迭代震荡及早熟收敛的情况。因此，研究人员尝试使用局部邻居的概念，将邻域也作为粒子进化的一个调节源，降低早熟收敛情况的发生概率。

在 PSO 算法中，粒子群的信息共享范围即为粒子的邻域拓扑结构。环形邻域拓扑结构使用局部邻居的概念，每个粒子只与最近的邻居沟通，较好地协调粒子本身和群体之间的关系。本文通过引入环形拓扑结构邻域改善 PSO 聚类算法性能。在初始阶段，邻域就是每个粒子自身，随迭代次数增加，每个粒子只与最近邻居沟通，邻域逐步扩展到包含所有粒子[11]。新的速度更新策略调整为：

$$v_i = \omega v_i + c_1 r_1(p_i - s_i) + c_2 r_2(g - s_i) + c_3 r_3(p_l - s_i) \tag{9}$$

其中，p_1 为粒子邻域极值。

3.2 改进的粒子群优化模糊 C 均值聚类

综上分析，本文提出的 IFCM-PSO 算法将聚类中心作为种群中粒子的位置，将 FCM 算法目标函数作为适应函数，终止条件为最优粒子目标函数适应值变化量小于阈值或迭代次数达到设定值 $iter_{totle}$，算法归纳如下：

(1) 设定聚类初始参数：聚类数，种群数，最大速度系数，迭代误差。

(2) 在取值范围 $[X_{min}, X_{max}]$ 内初始化聚类中心矩阵，并赋值给各粒子。

(3) 根据式(1) 计算初始种群中每个个体的适应值。

(4) 根据公式(9) 计算粒子移动速度，根据公式(6) 更新粒子的位置。

(5) 计算种群中个体粒子的适应值，若满足终止条件，则将粒子全局最优位置作为最优解输出；否则返回步骤(3) 继续迭代计算。

4 实验与结果分析

为了验证算法的性能，选择来自机器学习数据库 UCI 中的 3 个真实数据集进行实验，分别为 IRIS、WINE 和 Breast Cancer Wisconsin(BCW)。以上 3 个数据集经常被用于测试聚类算法的有效性，数据集的详细信息如表 1 所示。

表 1 数据集信息

数据集	样本总数	样本维数	分类数	样本分布
IRIS	150	4	3	50;50;50
WINE	178	13	3	59;71;48
BCW	699	9	2	458;241

4.1 算法有效性测试

对选择的 3 个数据集分别采用 FCM 算法、FCM-PSO 算法以及本文的 IFCM-PSO 算法进行聚类仿真实验。实验参数为：FCM-PSO 算法的粒子种群数为 20，最大迭代次数为 500，最优解改变量阈值为 0.001；IFCM-PSO 算法的粒子种群数为 20，允许的最大速度系数 $\lambda = 0.15$，最大迭代次数为 100，最优解改变量阈值为 0.001。数据集分别对 3 种算法进行 10 次仿真运算，各指标为 10 次计算的平均值，聚类结果如表 2 所示。

表 2 数据集聚类结果

数据集	算法	目标函数值	迭代次数	正确率/%
IRIS	FCM	60.58	28	89
	FCM-PSO	59.44	500	91
	IFCM-PSO	58.33	100	93
WINE	FCM	1.79×10^6	54	68
	FCM-PSO	2.04×10^7	500	58
	IFCM-PSO	1.27×10^6	100	76
BCW	FCM	1.52×10^4	14	94
	FCM-PSO	1.25×10^4	500	96
	IFCM-PSO	1.26×10^4	100	96

由表 2 可知，对 3 个数据集，FCM 算法迭代次数最少，表明收敛最快，但由于自身算法的缺陷使得聚类准确率较差；FCM-PSO 算法对 IRIS 和 BCW 两个数据集的聚类准确率较 FCM 算法高，但在 3 种算法中迭代次数最多，收敛速度最慢；本文的 IFCM-PSO 算法对 3 个数据集在迭代 100 次后均获得了最高的准确率，表明该算法在聚类速度和准确率方面的综合性能最好。

4.2 算法结果分析

对应 3 个数据集，FCM 算法、FCM-PSO 算法和 IF-

CM-PSO 算法各选取与聚类结果平均值最接近的一次聚类运算目标函数迭代曲线进行分析，目标函数值迭代曲线如图 1 所示。

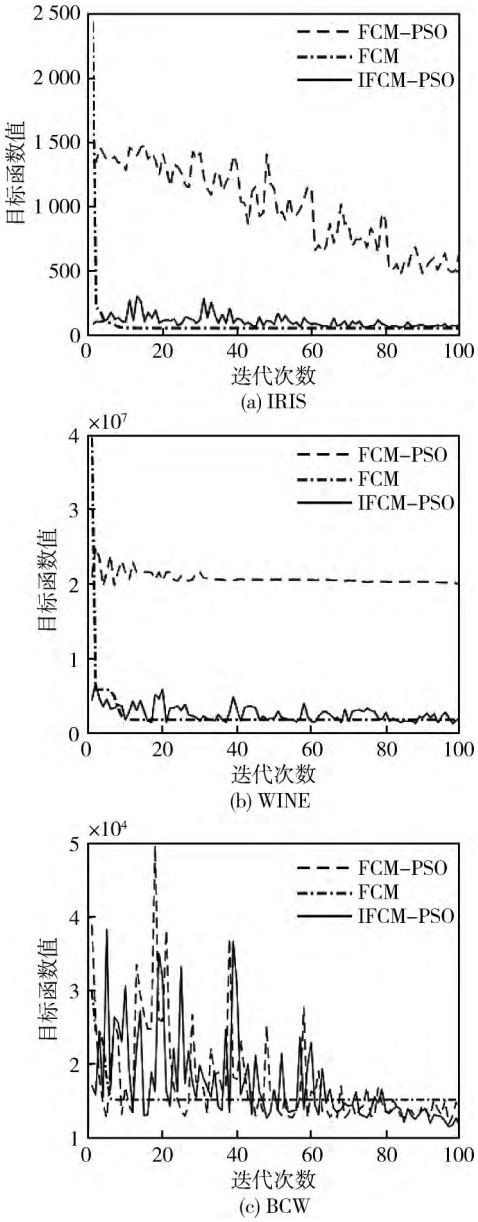

图 1　目标函数值迭代曲线图

由图 1(a) 可以发现，对 IRIS 数据集聚类时，FCM 算法函数值下降迅速，很快收敛; FCM-PSO 算法目标函数值在迭代 100 次后仍震荡，未见明显收敛; 而 IFCM-PSO 算法由于初始化取值接近最优解，收敛较快，目标函数值最小。

图 1(b) 显示，对 WINE 数据集，FCM 算法很快收敛，FCM-PSO 算法迭代约 30 次后收敛，但目标函数未见明显下降，表明出现早熟收敛; IFCM-PSO 算法在迭代 100 次后基本收敛，目标函数值与 FCM 算法目标函数值接近。

图 1(c) 显示对 Breast Cancer Wisconsin 数据集虽然 FCM-PSO 算法和本文的 IFCM-PSO 算法均出现震荡，但最终本文的 IFCM-PSO 算法震荡幅度较小，收敛效果更好。

通过以上 3 种算法对应 3 个数据集的目标函数曲线比较可以发现: 本文的 IFCM-PSO 聚类算法由于在聚类初始化取值、最大速度取值方面进行了改进，并引入了环形邻域辅助进化，使该算法有效克服了 FCM 算法对初始值敏感、易陷入局部最优解及基本粒子群聚类算法迭代初期震荡、早熟收敛的问题，因而获得了最好的聚类效果。

5　结束语

本文针对模糊 C 均值聚类算法存在的主要问题，利用改进的粒子群聚类算法，提出了一种基于粒子群优化的模糊 C 均值聚类算法。通过对粒子初始化空间和粒子运动最大速度两个参数的优化设置，并引入环形拓扑结构的邻域，提高了粒子群聚类算法的聚类效果。仿真结果表明该算法在聚类准确性和收敛速度方面均优于模糊 C 均值聚类(FCM) 算法和基本粒子群聚类(FCM-PSO) 算法。

参考文献

[1] 贺正洪，雷英杰. 直觉模糊 C 均值聚类算法研究 [J]. 控制与决策，2011，26(6) : 847-850，856.

[2] PIMENTEL B A，SOUZA R M. A weighted multivariate fuzzy C-means method in interval-valued scientific production data [J]. Expert Systems with Applications，2014，41(7) ，3223-3236.

[3] 杨慧，吴沛泽，倪继良. 基于改进粒子群置信规则库参数训练算法 [J]. 计算机工程与设计，2017，38(2) : 400-404.

[4] FARHAD S，AMIN A N，SHAHIN R N，et al. Evaluating the potential of particle swarm optimization in clustering of hyper-spectral imagery using fuzzy C-means [C] // International Conference on Asia Agriculture and Animal，Singapore: IACSIT，2011: 201-207.

[5] Niu Qiang，Huang Xinjian. An improved fuzzy C-means clustering algorithm based on PSO [J]. Journal of Software，2011，6(5) : 873-879.

[6] 王纵虎，刘志镜，陈东辉. 基于粒子群优化的模糊 C - 均值聚类算法研究 [J]. 计算机科学，2012，39(9) : 166-169.

[7] KUMUTHA V，PALANIAMMAL S. Improved fuzzy clustering method based on intuitionistic fuzzy particle swarm optimization [J]. Journal of Theoretical and Applied Information Technology，2014，62 (1) : 8-15.

[8] Chen Shouwen，Xu Zhuoming，Tang Yan. A hybrid clustering algorithm based on fuzzy C-means and improved particle swarm optimization [J]. Arabian Journal for Science & Engineering，2014，39(12) : 8875-8887.

(下转第 44 页)

scale visual recognition challenge [J]. International Journal of Computer Vision, 2015, 115 (3): 211-252.
[6] HINTON G, DENG L, YU D, et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups [J]. IEEE Signal Processing Magazine, 2012,29 (6): 82-97.
[7] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets [C] //Advances in Neural Information Processing Systems, 2014: 2672-2680.
[8] MIRZA M, OSINDERO S. Conditional generative adversarial nets [J]. arXiv preprint arXiv: 1411.1784. 2014.
[9] ISOLA P, ZHU J Y, ZHOU T, et al. Image-to-image translation with conditional adversarial networks [J]. arXiv preprint arXiv: 1611.07004. 2016.
[10] PATHAK D, KRAHENBUHL P, DONAHUE J, et al. Context encoders: feature learning by in painting [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 2536-2544.
[11] WANG S, LIU Z, LV S, et al. A natural visible and infrared facial expression database for expression recognition and emotion inference [J]. IEEE Transactions on Multimedia, 2010, 12(7): 682-691.
[12] BREULEUX O, BENGIO Y, VINCENT P. Quickly generating representative samples from an rbm derived process [J]. Neural Computation, 2011, 23 (8): 2058-2073.

(收稿日期: 2018-04-14)

作者简介:
王雅欣(1991 -),女,硕士研究生,主要研究方向: 情感计算。
史潇潇(1991 -),女,硕士,主要研究方向: 情感计算。

(上接第 39 页)

[9] 李峻金,向阳,芦英明,等. 粒子群聚类算法综述 [J]. 计算机应用研究, 2009,26(12): 4423-4427.
[10] FILHO T M S, PIMENTEL B A, SOUZA R M C R, et al. Hybrid methods for fuzzy clustering based on fuzzy C-means and improved particle swarm optimization [J]. Expert Systems with Applications, 2015, 42(17): 6315-6328.
[11] 石松,陈云. 层次环形拓扑结构的动态粒子群算法 [J]. 计算机工程与应用, 2013, 49(8): 1-5.

(收稿日期: 2018-04-29)

作者简介:
王宇钢(1977 -),男,博士研究生,讲师,主要研究方向: 机械制造自动化。

电子六所承办首个国家级“工控信息安全培训”

近日,由中国电子旗下电子六所承担的首个国家级“工控信息安全培训”成功举办。本次培训采取“线上 + 面授”相结合的方式,电子六所工控信息安全领域专家针对学员行业特征,讲授国内国际形势、相关法律法规、技术应用与系统解决方案等,并重点分享电力企业工控防护知识。