优化的核模糊C均值聚类算法

刘奕麟，安建成
（太原理工大学 计算机科学与技术学院，山西 太原 030024）

摘 要：提出一种优化的核模糊C均值聚类算法（WBAKFCM）.该算法首先通过改进蝙蝠算法（Weight bat Algorithm，WBA）确定最优聚类中心集合，然后用核模糊C均值聚类算法指导聚类划分.一方面，改进的蝙蝠算法在传统的蝙蝠算法中引入佳点集理论和速度权重，分别用于调节种群的初始化和个体位置的自适应更新.另一方面，在核模糊C均值聚类算法（Kernel Fuzzy C-Means，KFCM）中，选用了高斯核函数，从而将数据映射到高维特征空间进行聚类划分.实验结果表明，优化的核模糊C均值聚类算法在聚类准确率与时间效率上明显优于传统算法.

关键词：模糊C均值聚类；核函数；蝙蝠算法；佳点集；速度权重
中图分类号：TP391 **文献标识码**：A **文章编号**：1000－7180(2018)02－0079－05

Optimized Kernel Fuzzy C-Means Clustering Algorithm

LIU Yi-lin，AN Jian-cheng
(College of Computer Science and Technology，Taiyuan University of Technology，Taiyuan 030024，China)

Abstract：This paper an optimized kernel fuzzy C-means clustering algorithm (WBAKFCM) is proposed. Firstly，the optimal clustering center is found by the improved bat algorithm (WBA)，then the Kernel Fuzzy C-Means clustering algorithm (KFCM) is used to guide the clustering. On the one hand，the improved bat algorithm adding two strategies to the traditional bat algorithm，the good point set theory and velocity weight are used to adjust population initialization and adaptive updates of the individual position respectively. On the other hand，in the Kernel Fuzzy C-Means clustering algorithm，the Gaussian kernel function is selected to map the data to high-dimensional feature space for clustering. The experimental results show that the optimized kernel fuzzy C-means clustering algorithm is superior to the traditional algorithm in clustering accuracy and time efficiency.

Key words：Fuzzy C-Means；kernel function；Bat Algorithm；good point set；speed weight

1 引言

核模糊C均值聚类算法（KFCM）是在传统的模糊C均值聚类算法（Fuzzy C-Means，FCM）中引入核（kernel）的思想，解决线性不可分的问题，实现对各种数据结构的有效聚类[1]，目前已广泛应用多种领域.但其仍然存在很多问题，如对初始聚类中心敏感、计算时间长[1]、核函数和参数选择的多样性等.

针对以上问题，国内外相关学者提出了众多改进的方法.L Chen等[2]提出了广义多核的核模糊C均值聚类算法，该算法可以根据图像中的不同像素信息灵活的选择核函数，提供了一个统一的框架；Y Ding等[3]将遗传算法（Genetic Algorithm，GA）应用到KFCM中.

对此，本文提出了一种基于改进蝙蝠算法的核模糊C均值聚类算法（WBAKFCM）.

2 核模糊C均值聚类算法

2.1 FCM算法简述

FCM算法[4]是基于目标函数的模糊聚类算法，其核心思想是：数据集 $X=\{x_1,x_2,\cdots,x_n\}$ 中的各

收稿日期：2017－04－23；修回日期：2017－06－02
基金项目：山西省国际科技合作项目（2014081018-2）

样本 $x_i(i=1,2,\cdots,n)$ 对于聚类中心集合 $C=\{c_1,c_2,\cdots,c_k\}$ 中的各中心都有一个隶属度 $u_{ij}(i=1,2,\cdots,n;j=1,2,\cdots,k)$. 而且这也是与传统硬聚类算法如 K-means 聚类的本质区别,在传统硬聚类算法中,样本只能是属于或者不属于某个簇. 模糊 C 均值聚类算法引入模糊理论的思想,使目标函数最小,从而找到最优的聚类中心与隶属度. FCM 的表达形式为

$$J_{FCM}(\boldsymbol{U},C)=\sum_{i=1}^{n}\sum_{j=1}^{k}u_{ij}{}^{m}\parallel x_i-c_j\parallel^{2} \tag{1}$$

式中,m 表示模糊度,一般取值为 2,其值越大,表示聚类越模糊;$\parallel x_i-c_j\parallel^{2}$ 为样本到聚类中心的距离;隶属度矩阵 $\boldsymbol{U}$ 中的元素 u_{ij} 满足约束条件:

$$u_{ij}\in[0,1];\sum_{i=1}^{n}u_{ij}>0;\sum_{j=1}^{k}u_{ij}=1 \tag{2}$$

式(2)表明了隶属度的取值范围,且对于每个聚类中心,至少有一个数据样本属于该聚类中心所代表的簇,每个样本对于所有聚类中心的隶属度和为 1.

2.2 KFCM 算法

KFCM 算法在传统 FCM 算法中引入核的思想,用以解决数据线性不可分的问题[1]. 核是一个函数[5],对所有 $x,c\in\chi$,满足 $K(x,c)=\langle\varphi(x),\varphi(c)\rangle$,$\varphi$ 是从空间 χ 到特征空间 ρ 的映射. 低维特征空间中线性不可分的样本点,映射到高维特征空间,将变成线性可分的. 但采用直接映射的方式易引起维度爆炸. 而核函数是在原本的低维特征空间进行计算,等价于在高维特征空间的内积计算. KFCM 算法的目标函数为

$$J_{KFCM}(U,C)=\sum_{i=1}^{n}\sum_{j=1}^{k}u_{ij}^{m}\parallel\varphi(x_i)-\varphi(c_j)\parallel^{2}=\sum_{i=1}^{n}\sum_{j=1}^{k}u_{ij}^{m}(K(x_i,x_i)-2K(x_i,c_j)+K(c_j,c_j)) \tag{3}$$

对于一个映射,构造相应的核函数是很困难的,一般从常用的核函数中选择,如多项式核函数、线性核函数、高斯核函数. 由于高斯核函数对应的特征空间可以是无限维的,数据样本在该特征空间一定是线性可分的,因此本文选用高斯核函数,表达式如式(4)所示.

$$K(x,y)=e^{-\frac{\parallel x-y\parallel^{2}}{2a^{2}}} \tag{4}$$

引入高斯核函数后,KFCM 算法的目标函数简化为

$$J_{KFCM}(U,C)=2\sum_{i=1}^{n}\sum_{j=1}^{k}u_{ij}^{m}(1-K(x_i,c_j)) \tag{5}$$

约束条件与式(2)相同. 通过拉格朗日乘数法可得隶属度的更新公式为

$$u_{ij}=\frac{(1-K(x_i,c_j))^{-\frac{1}{m-1}}}{\sum_{j=1}^{k}(1-K(x_i,c_j))^{-\frac{1}{m-1}}} \tag{6}$$

聚类中心的更新公式为

$$c_j=\frac{\sum_{i=1}^{n}u_{ij}^{m}K(x_i,c_j)x_i}{\sum_{i=1}^{n}u_{ij}^{m}K(x_i,c_j)} \tag{7}$$

算法根据式(6)、(7)不断迭代循环更新隶属度矩阵,直到找到最优聚类中心,然后去模糊化得出最终的聚类结果.

3 蝙蝠算法

3.1 蝙蝠算法

蝙蝠算法(Bat Algorithm, BA)是一种新型启发式群智能算法. 该算法与粒子群算法(Particle Swarm Optimization, PSO)相似,模仿种群觅食的过程. 蝙蝠算法根据蝙蝠搜寻猎物时的回声定位原理,建立速度位移模型,并且在搜索猎物的过程中动态的调整脉冲频率、频度、响度,通过模拟这一过程寻找到最佳位置(即全局最优解). 与新型的仿生算法,如萤火虫算法、蜂群算法等相比较,蝙蝠算法[6]具有较高的寻优性能和收敛率. 传统的蝙蝠算法需要随机初始化种群,在搜索后期蝙蝠个体进行局部搜索时耗时过长,这些问题都会影响算法的全局寻优性能.

3.2 改进的蝙蝠算法

本文从两个方面对传统蝙蝠算法进行改进:运用佳点集生成方法对蝙蝠群体进行初始化;在蝙蝠速度更新公式中引入速度权重.

(1)佳点集理论

群智能算法的种群初始化一般采用随机生成的方法,该方法无法保证初始种群的均匀分布,影响算法的寻优效果. 现阶段,混沌映射[7]及佳点集常被用于种群的初始化,且文献[8]已证明随机生成、混沌映射、佳点集这三种初始化方式中,运用佳点集生成方法初始化的样本点分布更加均匀. 所以本文采用佳点集生成方法对蝙蝠群体进行初始化.

佳点集理论由数学家华罗庚、王元等在《数论在近似分析中的应用》中提出,基本定义如下:设 G 是 D 维空间中的单位立方体,令 $\langle\langle r\rangle\rangle\in G$ 形为

$P_{n(i)}=\{(r_1\times i,r_2\times i,\cdots,r_j\times i),i=1,2,\cdots,n\}$ 的偏差 $Q_{(n)}$ 满足 $Q_{(n)}=C_{(r,x)}n^{-1+x}$，其中 $C_{(r,x)}$ 是只与〈〈r〉〉、x 有关的常数，则称 $P_{n(i)}$ 为佳点集.

(2) 速度权重

传统蝙蝠算法中，蝙蝠的游走主要是由历史参照和与全局最优位置的差异决定的，此模式可以保证蝙蝠搜索时发散的广度，但在算法执行后期，个体容易偏离最优位置，不利于局部的搜索. 因此引入速度权重[9]：

$$v_i^{t+1}=\omega v_i{}^t+(x_i{}^t-x^*)f_i(0<\omega\leqslant 1) \tag{8}$$

蝙蝠进行全局搜索时采用较大的权重，保证搜索的广度；进行局部搜索时采用较小的权重，保证在局部范围内游走. 因此引入速度权重后可以协调算法的全局搜索与局部搜索.

本文提出的权重定义方法可以使权重根据搜索模式的类别进行自适应的变化：

$$\omega=e^{p^t-1} \tag{9}$$

$$s_i^t=\begin{cases}1,\text{if}(f(C_i)t>f(C_i)^{t-1})\\0,\text{other}\end{cases},p^t=\frac{\sum_{i=1}^{h}S_i{}^t}{h} \tag{10}$$

式中，当蝙蝠个体 i 的适应度值 $f(C_i)^t>f(C_i)^{t-1}$ 时，表示蝙蝠需要进一步搜索，$S_i{}^t$ 的取值为 1，否则为 0；p^t 表示 t 时刻蝙蝠群体需要进一步搜索的概率，且有 $0\leqslant p^t\leqslant 1$. 当 p^t 较大时，权重 ω 也较大；当 p^t 较小时，ω 迅速减小. p^t 与权重 ω 的函数关系呈指数型，如式(9) 所示.

改进的蝙蝠算法流程如下所示.

步骤 1：设置蝙蝠群体规模 h，最大迭代次数 T 等相关参数；

步骤 2：初始化蝙蝠种群，个体速度 v_i^0、位置 x_i^0、脉冲响度 A_i^0、脉冲频度 r_i^0、脉冲频率 f_i；

步骤 3：定义适应度函数；

步骤 4：迭代循环

While ($t<T$)

{

计算蝙蝠个体的适应度值，找出当前最佳解；

根据公式(9)、(10) 计算速度权重 ω；

根据 $f_i=f_{\min}+(f_{\max}-f_{\min})\beta$ 调整脉冲频率；

根据 $x_i^{t+1}=x_i^t+v_i^{t+1}$，以及式(8) 更新蝙蝠的速度、位置；

对每个蝙蝠个体产生一个随机数 rand1；

If (rand1 $>r_i$)

{

根据 $x_{i\text{new}}^t=x^*+\lambda\overline{A^i}$ 在当前最佳解附近形成一个新解，计算新的适应度值；

}

对每个蝙蝠产生一个随机数 rand2；

If (rand2 $<A_i$ 且 $f(x_{\text{new}})<f(x_i)$)

{

接受这个新的解；

根据 $r_i^t=r_i^0(1-e^{-\gamma t})$ 增大脉冲频度 r_i，根据 $A_i^{t+1}=\theta A_i^t$ 减小脉冲响度 A_i；

}

找到当前最佳解 x^*；

}

步骤 5：显示最后得到的最佳解.

4　优化的核模糊 C 均值聚类算法

核模糊 C 均值聚类算法需要随机初始化聚类中心，对于不同的数据样本需要人为定义聚类个数，造成聚类结果具有很强的差异性[10]. 对此，提出一种基于改进蝙蝠算法的核模糊 C 均值聚类算法，即用寻优性能好、收敛率高的改进蝙蝠算法先确定最优聚类中心集合，然后用核模糊 C 均值聚类算法指导聚类划分. KFCM 算法是寻求最小的目标函数值来确定最优聚类中心，而蝙蝠算法达到最优位置时适应度函数值最小，即定义蝙蝠算法的适应度函数为

$$f(C)=J_{KFCM}(U,C) \tag{11}$$

式中，$J_{KFCM}(U,C)$ 根据式(5) 计算.

若数据样本集为 $X=\{x_1,x_2,\cdots,x_n\}$，其元素 x_i 为任意维数的向量，聚类中心集合为 $C_i=\{c_1,c_2,\cdots,c_k\}$，蝙蝠算法中的每只蝙蝠代表一种聚类中心的集合，则蝙蝠群体表示为 $C=\{C_1,C_2,\cdots,C_h\}$，优化后的核模糊 C 均值聚类算法 WBAKFCM 算法简要流程图如图 1 所示.

步骤 1：确定蝙蝠群体规模 h，模糊度 m，收敛精度 ε，核参数 α，最大迭代次数 T 等参数；

步骤 2：初始化蝙蝠群体 C，其蝙蝠个体 C_i 在各维度的位置表示聚类中心的集合，初始化蝙蝠个体的速度 $v_i{}^0$、位置 $x_i{}^0$、脉冲频率 f_i；

步骤 3：使用式(6) 计算隶属度 u_{ij}，得到隶属度矩阵 U；

步骤 4：根据式(11) 计算蝙蝠个体的适应度值，选择最优个体；

步骤 5：根据式(9) 和(10) 更新速度权重；

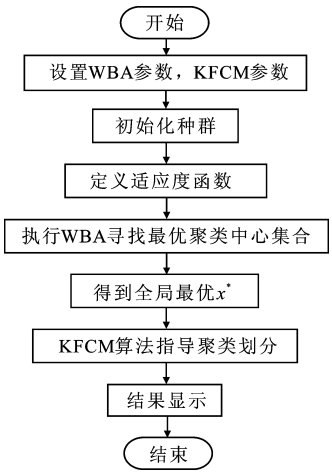

图 1 WBAKFCM 算法流程

步骤 6：根据式(8) 更新蝙蝠个体的速度；

步骤 7：根据式 $x_i^{t+1} = x_i^t + v_i^{t+1}$ 更新蝙蝠个体的位置；

步骤 8：生成随机数 rand1，若 rand1 $> r_i$，根据 $x_{inew}^t = x^* + \lambda \overline{A^t}$ 在当前最优位置附近进行搜索，形成一个新的解；

步骤 9：计算更新位置后蝙蝠个体的适应度值；

步骤 10：生成随机数 rand2，若 rand2 $< A_i$ 且适应度值 $f(x_{new}) < f(x_i)$，接受这个新解，调整脉冲响度；

步骤 11：更新最优解 x^*；

步骤 12：根据终止条件 $\max \| U^t - U^{t-1} \| < \varepsilon$ 与最大迭代次数 $t < T$ 判断是否终止，否则转到步骤 3.

步骤 13：得到最优聚类中心集合

步骤 14：进行聚类划分，显示结果.

5 实验结果

本文所有实验的运行环境均为 Window7 系统下 MATLAB R2015b，处理器 Intel(R) Core(TM) i5-3210M CPU @ 2.50 GHz，内存为 4 GB.

实验 1 选取 Benchmarks 测试函数集中具有代表性的 Sphere 单峰函数和 Rastrigin 多峰函数测试改进蝙蝠算法(WBA)的寻优性能. 对于传统的蝙蝠算法(BA)与改进的蝙蝠算法(WBA)，分别独立运行 30 次. 参数设置：蝙蝠群体规模 $h=20$，搜索维度 $d=10$，最大进化代数 1 000，脉冲频率范围 $[0,2]$，脉冲响度衰减系数 $\theta=0.9$，脉冲频度增强系数 $\eta=0.9$. 测试函数如表 1 所示. 表 2 为两种算法在不同测试函数下适应度的最优值、平均值.

表 1 测试函数

函数名称	函数表达式	函数全局最小值
Sphere	$f_1(x)=\sum_{i=1}^{d} x_i 2$	$f_{\min}(0,0,\cdots 0)=0$
Rastrigin	$f_2(x)=\sum_{i=1}^{d}(x_i 2-100\cos(2\pi x_i)+10)$	$f_{\min}(0,0,\cdots 0)=0$

表 2 BA 与 WBA 算法对比

函数	寻优值	BA	WBA
f_1	最优值	1.809e-07	1.3295e-10
	平均值	1.3087e-06	8.9467e-09
f_2	最优值	3.406e-08	2.5988e-11
	平均值	2.1262e-07	2.2509e-11

由表 2 中的实验数据可得，改进后的蝙蝠算法与传统蝙蝠算法相比，其全局最优值更加接近函数实际最小值，寻优能力更好.

此外，比较传统蝙蝠算法与改进蝙蝠算法在 Sphere 函数和 Rastrigin 函数上寻优时的收敛性能. 从图 2、3 可以发现改进后的蝙蝠算法达到全局最优时所需要的迭代次数明显小于传统算法，可以更快的找到最优值，所以改进蝙蝠算法的收敛性更好.

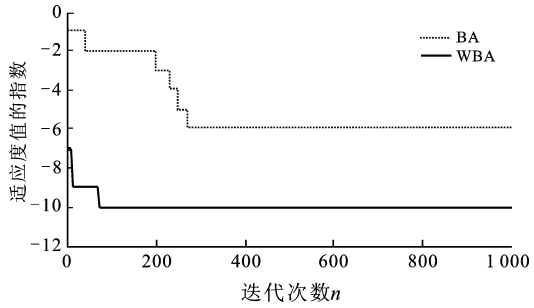

图 2 Sphere 函数收敛曲线

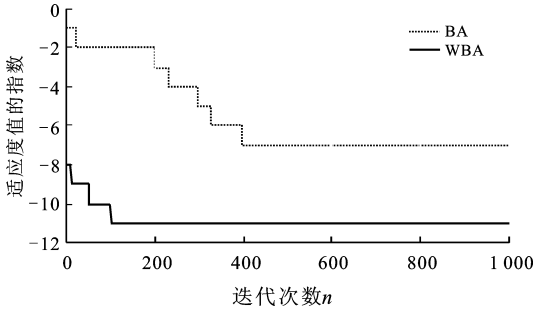

图 3 Rastrigin 函数收敛曲线

实验 2 采用 UCI 数据库中的 Wine、IRIS 公共数据集，测试本文提出的 WBAKFCM 算法的聚类准确率及时间效率. 数据集描述如表 3 所示.

WBAKFCM 算法参数设置：模糊度 $m=2$，收敛精度 $\varepsilon=10e-5$，对于两种数据集分别用 FCM、KFCM、文献[3]中的 GAKFCM、WBAKFCM 四种聚类算法，就聚类准确率与运行时间两方面进行实验对比分析. 实验结果如表 4、表 5 所示.

表 3　公共数据集

数据集	类别数	样本数	特征维数
Wine	3	178	13
IRIS	3	150	4

表 4　四种算法的聚类准确率对比　%

数据集	FCM	KFCM	GAKFCM	WBAKFCM
Wine	68.5	75.3	82	90.5
IRIS	89.3	92	93	96

表 5　四种算法的运行时间对比　s

数据集	FCM	KFCM	GAKFCM	WBAKFCM
Wine	22	26	33	27
IRI	11	22.5	32	24.7

实验数据表明，本文提出的 WBAKFCM 算法在聚类准确率上明显高于 FCM、KFCM、GAKFCM 算法，达到了较好的聚类效果；在运行时间上，由于算法加入核函数和蝙蝠算法的寻优迭代，时间耗费略高于 FCM、KFCM 算法，但是比 GAKFCM 算法达到聚类的时间减少了近 7 s. 综上，WBAKFCM 算法在达到较高的聚类准确率的同时，时间耗费相对减少，对于复杂样本以及实时的应用具有较高的优势及积极的参考作用.

6　结束语

本文对提出的优化的核模糊 C 均值聚类算法进行了详细描述，即通过用改进的蝙蝠算法寻找到最优聚类中心集合，然后根据核模糊 C 均值聚类算法进行聚类划分. 并且通过实验证明改进蝙蝠算法具有较好的寻优性能，就本文提出的优化算法 WBAKFCM 与 FCM 算法、KFCM 算法、GAKFCM 算法分别在 Wine 及 IRIS 数据集上进行聚类，实验结果表明 WBAKFCM 算法在达到较高的聚类准确率的同时减少了时间的耗费. 本文算法的不足之处在于核参数的选取依靠经验值来确定，因此，如何根据不同的数据自适应的设置参数值，研究适应性更强的算法是今后的研究重点.

参考文献：

[1] 伍忠东，高新波，谢维信. 基于核方法的模糊聚类算法[J]. 西安电子科技大学学报(自然科学版)，2004，31(4)：533-537.

[2] Chen L，Chen C L，Lu M. A multiple-kernel fuzzy C-means algorithm for image segmentation. [J]. IEEE Transactions on Systems Man & Cybernetics Part B Cybernetics A Publication of the IEEE Systems Man & Cybernetics Society，2011，41(5)：1263.

[3] Ding Y，Fu X. Kernel-based fuzzy c-means clustering algorithm based on genetic algorithm[J]. Neurocomputing，2015(188)：233-238.

[4] Bezdek J，Hathaway R，Sobin M，et al. Convergence theory for fuzzy c-means：Counterexamples and repairs [J]. Systems Man & Cybernetics IEEE Transactions on，1987，17(5)：873-877.

[5] Scholkopf B，Mika S，Burges C J C，et al. Input space versus feature space in kernel-based methods [J]. IEEE Transactions on Neural Networks，1970，10(5)：1000-1017.

[6] 杨雁莹，徐仙伟，曹霁. 基于仿生理论的新型优化算法综述[J]. 计算机仿真，2016，33(6)：233-237.

[7] 张晓琳，张冲，杨涛. 基于改进布谷鸟算法的火焰图像阈值分割算法[J]. 微电子学与计算机，2017，34(1)：66-70.

[8] 刘万军，杨笑，曲海成. 基于 SQP 局部搜索的蝙蝠优化算法[J]. 计算机工程与应用，2016，52(15)：183-189.

[9] 薛菲. 基于蝙蝠算法的启发式智能优化研究与应用[D]. 北京：北京工业大学，2016.

[10] 余晓东，雷英杰，岳韶华，等. 基于粒子群优化的直觉模糊核聚类算法研究[J]. 通信学报，2015，36(5)：74-80.

作者简介：

刘奕麟　女，(1991-)，硕士. 研究方向为人工智能和图像处理. E-mail：654969188@qq.com.

安建成　男，(1963-)，副教授. 研究方向为机器学习与图像处理.