文章编号：2095-2163(2021)07-0073-07 中图分类号：TP181 文献标志码：A

基于余弦相似性的自适应权重的改进 FCM 算法

胡建华，尹慧琳
（上海理工大学 理学院，上海 200093）

摘 要： 模糊 C-均值聚类算法(FCM)是一种经典的聚类算法，主要通过迭代更新隶属度和聚类中心来提高聚类的有效性。FCM 算法的性能主要通过类内紧性和类间分离性来评价，但其既依赖于初始聚类中心，也对噪声非常敏感。考虑到每个数据点和每个聚类中心对目标函数的不同重要性，本文提出了一种具有自适应权重的改进 FCM 聚类算法(Hybrid FCM)。主要贡献：将 2 个具有自适应指数 p 和 q 的自适应权向量 ψ 和 φ 引入 FCM 的目标函数，以体现不同数据点和聚类中心的重要性；为提高聚类性能，自适应指数 p、q 和模糊因子 m 采用粒子群优化算法(PSO)优化，新提出的聚类评价指标 AWCVI 作为 PSO 算法的适应度函数；迭代过程中利用余弦相似性对隶属度函数进行修正，提高算法的鲁棒性。实验表明，本文提出的算法能够有效地提高聚类效果。

关键词： 模糊 C 均值算法；自适应权重；余弦相似度；粒子群算法

Improved FCM algorithm with adaptive weights based on cosine similarity

HU Jianhua, YIN Huilin
(College of Science, University of Shanghai for Science and Technology, Shanghai 200093, China)

【Abstract】 As a classical clustering algorithm, fuzzy C-means clustering algorithm (FCM) improves the effectiveness of clustering by updating membership and clustering centers. The performance of FCM algorithm is mainly evaluated by intra cluster compactness and inter cluster separation. But FCM algorithm relies on the clustering centers and is very sensitive to noise. Considering the different importance of each data point and each cluster center, two adaptive weight vectors ψ and φ with adaptive exponents p and q are introduced into the objective function of FCM, and an improved FCM clustering algorithm with adaptive weights is proposed; At the same time, the new clustering evaluation index AWCVI is used to optimize the parameters p, q and the fuzzy factor m, which is determined by the particle swarm optimization algorithm (PSO); The cosine similarity is used to modify the membership function in the iterative process to improve the robustness of the algorithm. Experimental results show that the proposed algorithm can effectively improve the clustering effect.

【Key words】 Fuzzy C-means algorithm(FCM); adaptive weight; cosine similarity; Particle Swarm Optimization

0 引 言

模糊 C 均值聚类算法是一种经典的聚类方法，由 Dunn[1]在 1973 年提出，由于其简单、易实现而广泛应用于数据挖掘、模式识别、信号处理、图像分割等领域中[2-4]。然而，FCM 算法依赖初始聚类中心、对噪声非常敏感、且容易陷入局部最优。因此，许多改进的 FCM 算法也相继提出以克服这些问题。一般 FCM 算法需要将样本对每个类满足归一化条件，这样会导致算法对非平衡数据集中噪声点和离群值异常敏感，文献[5]提出一种改进的模糊隶属度函数的 FCM 聚类算法，能在聚类过程中不断对隶属度进行修正，从而消除噪声点，提高聚类的有效性；Zhou 等人[6]结合了图像的空间信息和 FCM 算法，提出了一种自适应空间信息模糊聚类算法用来提高图像分割的效果；文献[7]建立了一种基于特征的 2 型模糊聚类算法，将 2 型模糊集引入到聚类算法中，可以更灵活地处理由噪声环境引起的与隶属度概念相关的不确定性；将 FCM 算法与其他算法结合起来，能够解决 FCM 算法依赖初始值的问题，例如将 FCM 算法与粒子群算法结合在一起[8-9]，利用粒子群算法强大的全局寻优能力，能够使 FCM 算法取得相对更优的初始点，提升聚类效果；文献[10]将 FCM 算法与蚁群算法结合起来，克服 FCM 算法依赖初始值的缺点。考虑到噪声和样本分布不均衡，文献[11]中提出了一种具有自适应样本权重的 FCM 算法(AFCM-SP)，通过对每个样本点赋予权重来区分不同的样本点对聚类结果的影响，并且用

基金项目： 国家自然科学基金(61873169)。

作者简介： 胡建华(1978-)，女，博士，讲师，主要研究方向：人工智能、李代数；尹慧琳(1996-)，女，硕士研究生，主要研究方向：人工智能、李代数。
通讯作者： 胡建华 Email：hjh_2021@usst.edu.cn

收稿日期： 2021-04-23

改进的粒子群算法(PSO-SP)对新引入的自适应参数进行优化,从而一定程度上降低了噪声点的影响。

但是由于样本的最终聚类结果也受到了聚类中心的影响,本文通过对样本与中心同时赋予权重的方法来提高算法聚类效果,基于自适应权重,还提出了新的聚类评价指标对聚类结果进行评价,同时,聚类过程中的隶属度函数可能会因为随机的初始值而降低聚类效果,所以对迭代过程中的隶属度进行余弦修正增强算法的鲁棒性,实验表明这一想法的确能够有效提高算法的性能。考虑到每个数据点和每个聚类中心对目标函数的不同重要性,本文提出了一种具有自适应权重的改进 FCM 聚类算法(Hybrid FCM)。在新算法中,2 个具有自适应指数 p 和 q 的自适应权向量 $\boldsymbol{\psi}$ 和 $\boldsymbol{\varphi}$ 引入 FCM 的目标函数,以区分不同数据点和聚类中心在迭代过程中的不同重要性;为提高聚类性能,自适应指数 p、q 和模糊因子 m 采用粒子群优化算法(PSO)优化;相应地,新提出一种带权重的聚类评价指标 AWCVI 用来刻画类内紧致度和类间分离性,并作为 PSO 算法的适应度函数;为了提高算法的鲁棒性,迭代过程中利用余弦相似性对隶属度函数进行修正。通过在 5 个数据集上与另外 3 种算法对比的实验表明,本文提出的算法能够有效地提高聚类效果。

1 相关工作

1.1 经典 FCM 算法

模糊 C-均值聚类算法(FCM)是一种经典的聚类算法,主要通过迭代更新隶属度和聚类中心来提高聚类的有效性。从模型上看,FCM 算法是用于求解最小化问题,以 C 均值函数作为目标函数:

$$J(U,V)=\sum_{i=1}^{n}\sum_{j=1}^{c}\mu_{ji}^{m}\parallel x_i-v_j\parallel^2 \tag{1}$$

其中,x_i 表示第 i 个样本;令 $\boldsymbol{U}=(\mu_{ji})_{c\times n}$ 表示模糊隶属度矩阵,μ_{ji} 表示第 i 个样本属于第 j 类的隶属度,满足约束条件:

$$\mu_{ji}\in[0,1]\quad\sum_{j=1}^{c}\mu_{ji}=1,i=1,2\cdots,n \tag{2}$$

研究中,$V=\{v_1,v_2,\cdots,v_c\}$ 是所有聚类中心的集合,c_i 为第 i 个聚类中心;m 为模糊因子,一般取值为 2。FCM 的隶属度函数和聚类中心的更新迭代公式为:

$$\mu_{ji}=\frac{1}{\sum_{k=1}^{c}\left(\frac{\parallel x_i-v_j\parallel}{\parallel x_i-v_k\parallel}\right)^{\frac{2}{m-1}}} \tag{3}$$

$$v_j=\frac{\sum_{i=1}^{n}\mu_{ji}^{m}x_i}{\sum_{i=1}^{n}\mu_{ji}^{m}} \tag{4}$$

1.2 余弦相似性

余弦相似性,也叫余弦距离,是用向量空间中 2 个向量夹角的余弦值作为衡量 2 个个体之间的差异的大小度量。设 $\boldsymbol{M}$ 和 $\boldsymbol{N}$ 为 2 个样本向量,其计算公式为:

$$similarity=\cos(\theta)=\frac{\boldsymbol{M}\cdot\boldsymbol{N}}{\parallel\boldsymbol{M}\parallel\ \parallel\boldsymbol{N}\parallel}=\frac{\sum_{i=1}^{n}M_iN_i}{\sqrt{\sum_{i=1}^{n}(M_i)^2}\sqrt{\sum_{i=1}^{n}(N_i)^2}} \tag{5}$$

其中,分子为 $\boldsymbol{M}$,$\boldsymbol{N}$ 的向量内积,分母为向量 $\boldsymbol{M}$,$\boldsymbol{N}$ 的模的乘积。余弦值取值范围为 [-1,1],越接近 1,就说明 2 个向量越相似。余弦值注重从方向上区分样本的差异,而对绝对的数值不敏感,可修正数据度量标准不统一等问题。在数据挖掘领域,余弦相似性常用来衡量聚类的类内凝聚程度。

1.3 粒子群优化算法

粒子群优化(PSO)算法因为其具有搜索速度快、效率高、算法简单等优点,成为近年来广泛使用的优化算法之一[12-15],其传统模型为:

$$\boldsymbol{V}_i(t+1)=\omega\,\boldsymbol{V}_i(t)+c_1\,\boldsymbol{r}_1(\boldsymbol{P}_i(t)-\boldsymbol{X}_i(t))+c_2\,\boldsymbol{r}_2(\boldsymbol{P}_g(t)-\boldsymbol{X}_i(t)) \tag{6}$$

$$\boldsymbol{X}_i(t+1)=\boldsymbol{X}_i(t)+\boldsymbol{V}_i(t+1) \tag{7}$$

其中,$\boldsymbol{V}_i(t)$ 表示在搜索空间内第 i 个粒子在 t 时刻的速度;$\boldsymbol{X}_i(t)$ 表示在 t 时刻的位移;ω 是惯性权重;c_1,c_2 是加速系数,分别称为认知加速系数和社会加速系数,本文设置 $\omega=0.729$,$c_1=c_2=1.49$。在寻优过程中,$\boldsymbol{P}_i$ 和 $\boldsymbol{P}_g$ 分别代表第 i 个粒子的个体最佳位置与全局最佳位置。

2 混合自适应加权 FCM 算法

2.1 混合自适应加权 FCM 模型

FCM 是一种快速有效的聚类算法,但在噪声和样本分布不均衡的情况下,算法聚类结果不理想。文献[11]考虑到每个样本的重要性,提出了具有样本自适应权重的 FCM 算法(AFCM-SP),一定程度上减少了噪声干扰,提高算法性能。但在聚类过程中,聚类中心也起着非常重要的作用。观察式(1),可以发现经典 FCM 算法的目标函数可以写成:

$$J = 1 \cdot \sum_{j=1}^{c} \mu_{j1}^{m} \| x_1 - v_j \|^2 + 1 \cdot \sum_{j=1}^{c} \mu_{j2}^{m} \| x_2 - v_j \|^2 + \cdots + 1 \cdot \sum_{j=1}^{c} \mu_{jn}^{m} \| x_n - v_j \|^2 \quad (8)$$

将 $D_i = \sum_{j=1}^{c} \mu_{ji}^{m} \| x_i - v_j \|^2$ 看成数据点 x_i 对目标函数 J_c 的贡献，则式(8)表明每个数据点对于目标函数具有相同的重要性，显然这是不合实际情况的。本文同时考虑到不同样本和聚类中心对目标函数的不同影响程度，提出混合自适应加权 FCM 算法(Hybrid FCM)，其目标函数如下：

$$G_h = \frac{1}{2} \sum_{i=1}^{n} \psi_i^p \sum_{j=1}^{c} \mu_{ji}^{m} \| x_i - v_j \|^2 + \frac{1}{2} \sum_{j=1}^{c} \varphi_j^q \sum_{i=1}^{n} \mu_{ji}^{m} \| x_i - v_j \|^2 \quad (9)$$

其约束条件为：

$$\sum_{j=1}^{c} \mu_{ij} = 1 \quad \prod_{i=1}^{n} \psi_i = 1 \quad \prod_{j=1}^{c} \varphi_j = 1 \quad (10)$$

其中，$\mu_{ij} \in [0,1]$；$\boldsymbol{\psi} = (\psi_1, \psi_2, \cdots, \psi_n)$ 为样本权重向量；$\boldsymbol{\varphi} = (\varphi_1, \varphi_2, \cdots, \varphi_c)$ 为中心权重向量；$\psi_i > 0, \varphi_j > 0$。为了最小化 G_h，引入了拉格朗日函数：

$$L(\boldsymbol{\psi}, \boldsymbol{\varphi}, \boldsymbol{U}, \boldsymbol{V}) = \frac{1}{2}\Big(\sum_{i=1}^{n} \psi_i^p \sum_{j=1}^{c} \mu_{ji}^{m} \| x_i - v_j \|^2 \Big) + \frac{1}{2}\Big(\sum_{j=1}^{c} \varphi_j^q \sum_{i=1}^{n} \mu_{ji}^{m} \| x_i - v_j \|^2 \Big) + \alpha\Big(\prod_{i=1}^{n} \psi_i - 1 \Big) + \beta\Big(\prod_{j=1}^{c} \varphi_j - 1 \Big) + \sum_{i=1}^{n} \lambda_i \Big(\sum_{j=1}^{c} \mu_{ji} - 1 \Big)$$

对任意的 i,j，对 L 计算与 $\mu_{ji}, \psi_i, \varphi_j, v_j$ 相关的偏导数，并使其分别等于 0，结合约束条件(10) $\prod_{l=1, l\neq i}^{n} \psi_l = \frac{1}{\psi_i}$，$\prod_{l=1, l\neq j}^{c} \varphi_l = \frac{1}{\varphi_j}$，得到使目标函数(9)达到最小的充要条件：

$$\mu_{ji} = \left\{ \sum_{k=1}^{c} \frac{(\psi_i + \varphi_j) \| x_i - v_j \|^2}{(\psi_i + \varphi_j) \| x_i - v_k \|^2} \right\}^{\frac{1}{m-1}} \Bigg\}^{-1} \quad (11)$$

$$\psi_i = \left(\frac{\left[\prod_{i=1}^{n} \left(\sum_{j=1}^{c} \mu_{ji}^{m} \| x_i - v_j \|^2 \right) \right]^{\frac{1}{n}}}{\sum_{j=1}^{c} \mu_{ji}^{m} \| x_i - v_j \|^2} \right)^{\frac{1}{p}} \quad (12)$$

$$\varphi_j = \left(\frac{\left[\prod_{j=1}^{c} \left(\sum_{i=1}^{n} \mu_{ji}^{m} \| x_i - v_j \|^2 \right) \right]^{\frac{1}{c}}}{\sum_{i=1}^{n} \mu_{ji}^{m} \| x_i - v_j \|^2} \right)^{\frac{1}{q}} \quad (13)$$

$$v_j = \frac{\sum_{i=1}^{n} (\psi_i + \varphi_j) \mu_{ji}^{m} x_i}{\sum_{i=1}^{n} (\psi_i + \varphi_j) \mu_{ji}^{m}} \quad (14)$$

将式(11)、式(14)与式(3)、式(4)对比，Hybrid FCM 的隶属度和中心都因为自适应随权向量 $\boldsymbol{\psi}$、$\boldsymbol{\varphi}$ 的引进做了相应的调整，这有利于克服 FCM 算法对初始聚类中心的依赖和减低噪声的干扰。为了降低算法陷入局部最优的可能性，隶属度函数进一步用余弦相似度作为矫正因子来修正，计算公式为[16]：

$$u_{ji}^{k+1} = \frac{1}{2} u_{ji}^{k} + \frac{1}{2} \frac{x_i \cdot v_j}{\| x_i \| \| v_j \|} \quad (15)$$

同时，Hybrid FCM 中，自适应参数 p,q 和 m 一样是个超参数，用来控制函数的凸性和模糊度，恰当的取值对聚类结果起着至关重要的作用。

2.2　基于自适应权重的聚类有效性指标

聚类有效性指标(AWCVI)是用来衡量聚类效果的评价函数，XB 指标[13]是最广泛使用的聚类有效性指标之一，具体公式为：

$$CVI_{XB} = \frac{\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{c} \mu_{ji}^{m} \| x_i - v_j \|^2}{\min_{j\neq k} \| v_j - v_k \|^2} \quad (16)$$

CVI_{XB} 值越小，则表示具有良好的类内紧凑性和类间的分离性，说明聚类结果越好。此外，类间的最小距离反映了类间分离的尺度，而类内的平均距离体现了集群内的紧凑性尺度。本文结合新引入的自适应权重，提出新的聚类有效性指标 AWCVI 如下：

$$AWSPT = \min_{j\neq k} \| \varphi_j v_j - \varphi_k v_k \|^2 \quad (17)$$

$$AWCOMP = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{c} \psi_i \mu_{ji}^{m} \| x_i - v_j \|^2 \quad (18)$$

$$AWCVI = \frac{AWCOMP}{AWSPT} = \frac{\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{c} \psi_i \mu_{ji}^{m} \| x_i - v_j \|^2}{\min_{j\neq k} \| \varphi_j v_j - \varphi_k v_k \|^2} \quad (19)$$

作为比值函数，$AWCVI$ 的值越小，表示聚类效果越好。

2.3　混合自适应加权 FCM 算法的流程

Hybrid FCM 算法分为 2 部分。首先，为了更好的聚类效果，新引入的自适应参数 p,q 和模糊因子 m 通过经典的 PSO 算法优化，基于自适应权重的聚类有效性指标 AWCVI 作为 PSO 算法的适应度函数。其次，以式(9)为最小化的目标函数，通过有限次的迭代，得到最终的聚类结果。算法流程如下：

(1)在三维搜索空间内初始化粒子种群 $X_i=(p_i,q_i,m_i)$, $i=1,2,\cdots,N$,初始化隶属度矩阵 $\boldsymbol{U}$,样本权重向量 $\boldsymbol{\psi}$,中心权重向量 $\boldsymbol{\varphi}$,通过式(10)计算初始聚类中心 $\boldsymbol{V}$,设置收敛阈值 *eps*。

(2)初始化粒子的速度和位移。

(3)根据式(11)~式(15)迭代更新 $\boldsymbol{U},\boldsymbol{\psi},\boldsymbol{\varphi},\boldsymbol{V}$。

(4)根据式(19)计算每个粒子的适应度 $AWCVI(X_i(t))$,得到 t 时刻的 $\boldsymbol{P}_i$, $\boldsymbol{P}_g$。

(5)根据式(6)、式(7)更新粒子的速度和位移。

(6)当达到最大迭代次数时或者当 $|AWCVI(\boldsymbol{P}_g(k))-AWCVI(\boldsymbol{P}_g(k-1))|<eps$ 时,输出全局最优粒子 $\boldsymbol{P}_g=(m,p,q)$,否则返回到(3)。

同时,研究又给出基于余弦相似性的自适应权重的改进 FCM 算法流程如下:

(7)初始化 $\boldsymbol{U}$, $\boldsymbol{\psi}$, $\boldsymbol{\varphi}$, 通过式(14)计算聚类中心 $\boldsymbol{V}$。

(8)根据式(11)~(14)迭代更新 $\boldsymbol{U}$, $\boldsymbol{\psi},\boldsymbol{\varphi},\boldsymbol{V}$。

(9)通过式(15)对更新后的 $\boldsymbol{U}$ 进行余弦修正。

(10)利用式(9)计算目标函数 G_h。

(11)当算法收敛时或达到最大迭代次数时,进行(5);否则,回到(2)。

(12)得到最终聚类结果。

在此基础上,研发得到的算法伪代码如下。

算法 1 基于自适应权重的聚类有效性指标 *AWCVI* 的 *PSO* 算法

输入: $\boldsymbol{X}_i=(m_i,p_i,q_i)$

输出: 最优的 m, p, q

初始化粒子群 $\boldsymbol{X}=\boldsymbol{X}(m,p,q)$, 隶属度函数 $\boldsymbol{U}$, 权重向量 $\boldsymbol{\psi}$, $\boldsymbol{\varphi}$, 并且计算初始聚类中心 $\boldsymbol{V}$

初始化 $\boldsymbol{P}_i$, $\boldsymbol{P}_g$, 设置最大迭代次数 *Itermax*, 收敛域 *eps*

function FISTNESS FUNCTION = $AWCVI(X)$

for $k=1$:*itermax* do

for $i=1$:N\\ N 为种群数量

if $AWCVI(X_i(k))<AWCVI(P_i(k))$ *then* \\ $AWCVI(X_i(k))$ 是第 k 次迭代的 AWCVI 值

$P_i(k)=X_i(k)$

end if

$\boldsymbol{V}_i(t+1)=\omega\boldsymbol{V}_i(t)+c_1\boldsymbol{r}_1(\boldsymbol{P}_i(t)-\boldsymbol{X}_i(t))+c_2\boldsymbol{r}_2(\boldsymbol{P}_g(t)-\boldsymbol{X}_i(t))$

$\boldsymbol{X}_i(t+1)=\boldsymbol{X}_i(t)+\boldsymbol{V}_i(t+1)$

$i=i+1$

end for \\ 更新粒子的速度与位置

$P_g=\arg\min AWCVI(\boldsymbol{P}_i(k))$

if $|AWCVI(\boldsymbol{P}_g(k))-AWCVI(\boldsymbol{P}_g(k-1))|<eps$ then

return $\boldsymbol{P}_g=(m,p,q)$ \\ 全局最优粒子

else

$k:=k+1$

end if

end for

Return $\boldsymbol{P}_g=(m,p,q)$

end function

算法 2 混合自适应加权 FCM 算法

输入: $\boldsymbol{P}_g=(m,p,q)$ \\ 由算法 1 可以得到全局最优的 m, p ,q 值

输出: 聚类结果

初始化 隶属度函数 $\boldsymbol{U}$, 权重向量 $\boldsymbol{\psi}$, $\boldsymbol{\varphi}$, 并且计算初始聚类中心 $\boldsymbol{V}$, 聚类数目 c, 设置最大运算次数 *Itermax*, 收敛阈值 *eps*

function OBJECT FUNCTION = $G_h(\boldsymbol{U},\boldsymbol{\psi},\boldsymbol{\varphi},\boldsymbol{V})$

for $k=1$: *Itermax* do

for $i=1$:n \\ n 为数据的数量

for $j=1$:c

计算 $\mu_{ji},\psi_i,\varphi_j,v_j$

\\更新隶属度, 权重向量和聚类中心

$$u_{ji}^{k+1}=\frac{1}{2}u_{ji}^{k}+\frac{1}{2}\frac{x_i\cdot v_j}{\|x_i\|\,\|v_j\|}$$

\\ 对隶属度进行余弦修正

end for

end for

if $|G_h(k)-G_h(k-1)|<eps$ then

return $G_h(k)$ \\ 算法收敛时的目标函数值

else

$k:=k+1$

end if

end for

Return 隶属度函数 $\boldsymbol{U}$,权重向量 $\boldsymbol{\psi}$, $\boldsymbol{\varphi}$, 聚类中心 $\boldsymbol{V}$, 与目标函数值 G_h

end function

3 仿真实验与结果分析

为了验证提出算法的聚类性能,本文对 5 个数据集进行仿真实验,人工数据集 Twomoons 用来验证算法的鲁棒性,数据集信息包括数据的数量、特征、类别,见表 1;同时采用 FCM 算法[1],AFCM-SP[11] 算法与 SFCM 算法[16] 作为对比算法。实验之初,随

机分配[0,1]之间的值给 u_{ji}，ψ_i 和 φ_j 设置为 1；在聚类过程的最大迭代次数为 200，收敛阈值 *eps* 为 10^{-5}；在优化参数 p,q,m 的 PSO 中，搜索空间为 3 维，种群粒子数为 20，最大迭代次数设置为 20。在 5 个数据集中由 PSO 算法得到的最优 p,q,m 值列在表中，详见表 2。

表 1　6 个数据集的分布特征

Tab. 1　Distribution characteristics of six datasets

数据集	样本数目	特征数目	类别数目
IRIS	150	4	2
SONAR	208	60	2
SYM	350	6	3
Twomoons	1 502	2	2
Spiral	567	2	2

表 2　由 PSO 算法确定的 p，q，m 值

Tab. 2　p，q，m values by PSO algorithm

数据集	p	q	m
IRIS	1.57	1.23	2.97
SONAR	3.60	4.02	1.65
SYM	1.40	2.30	1.98
SPIRAL	2.37	2.12	1.26
Twomoons	2.92	1.87	3.11

对于前四个数据集，本文从 Xie Beni 指标（CVI_{XB}）、聚类准确率（*Accuracy*）、标准互信息（*NMI*）三个方面对 4 种算法进行对比实验。CVI_{XB} 值越小、*Accuracy* 和 *NMI* 值越大说明聚类效果越好。实验结果见表 3~表 6。从表 3~表 6 中可以看出，本文提出的改进算法 Hybrid FCM 在 IRIS 上的各个指标都优于其它 3 种算法，在 SONAR 中，Hybrid FCM 和 AFCM-SP 算法有相同的准确率，但是 CVI_{XB} 和 *NMI* 值都优于 AFCM-SP；在 SPIRAL 中，经典的 FCM 有着较好的 CVI_{XB} 值，但是准确率和 *NMI* 值还是 Hybrid FCM 更为优异；在 SYM 上，Hybrid FCM 准确率略低于 SFCM 算法，但 CVI_{XB} 和 *NMI* 值都排在第一位。对于人工数据集 Twomoons，本文添加 3 个评价指标，即：召回率（*Recall*）、精确率（*Precision*）和 *F*1 值，结果见表 7。Hybrid FCM 在 *Accuracy*、*NMI*、*Recall*、F_1 值在 4 个方面都有优势；Twomoons 数据集经过 4 种算法聚类后的数据分布如图 1 所示，可以发现在类边界部分，改进的算法有着更好的效果。图 2 给出 4 种算法在不同数据集上的收敛曲线以证明算法良好的收敛性。综上所述，本文提出的 Hybrid FCM 算法在 5 个数据集上都体现出较为优异的性能，能有效提高聚类效果。

表 3　在 IRIS 上的 CVI_{XB}，*Accuracy*，*NMI*

Tab. 3　CVI_{XB}，*Accuracy*，*NMI* on IRIS

算法	CVI_{XB}	*Accuracy*	*NMI*
FCM	0.136 9	0.893 3	0.746 5
AFCM-SP	0.176 5	0.926 7	0.787 3
SFCM	1.040 2	0.886 7	0.737 5
Hybrid FCM	**0.024 1**	**0.930 3**	**0.789 3**

表 4　在 SONAR 上的 CVI_{XB}，*Accuracy*，*NMI*

Tab. 4　CVI_{XB}，*Accuracy*，*NMI* on SONAR

算法	CVI_{XB}	*Accuracy*	*NMI*
FCM	2.187 6	0.552 9	0.008 8
AFCM-SP	1.159 0	**0.562 5**	0.013 4
SFCM	4.171 1	0.548 1	0.007 2
Hybrid FCM	**1.155 0**	**0.562 5**	**0.014 4**

表 5　在 SPIRAL 上的 CVI_{XB}，*Accuracy*，*NMI*

Tab. 5　CVI_{XB}，*Accuracy*，*NMI* on SPIRAL

算法	CVI_{XB}	*Accuracy*	*NMI*
FCM	**0.279 3**	0.583 4	0.247 1
AFCM-SP	0.300 8	0.598 1	0.279 0
SFCM	0.569 0	0.594 5	0.265 4
Hybrid FCM	0.293 4	**0.600 3**	**0.279 3**

表 6　在 SYM 上的 CVI_{XB}，*Accuracy*，*NMI*

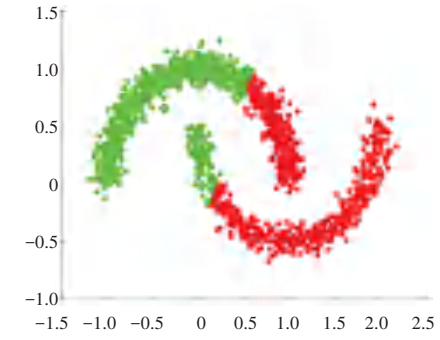
Tab. 6　CVI_{XB}，*Accuracy*，*NMI* on SYM

算法	CVI_{XB}	*Accuracy*	*NMI*
FCM	0.143 8	0.745 7	0.497 8
AFCM-SP	0.194 0	0.765 7	0.529 3
SFCM	0.942 7	**0.808 6**	0.556 9
Hybrid FCM	**0.115 9**	0.790 1	**0.577 8**

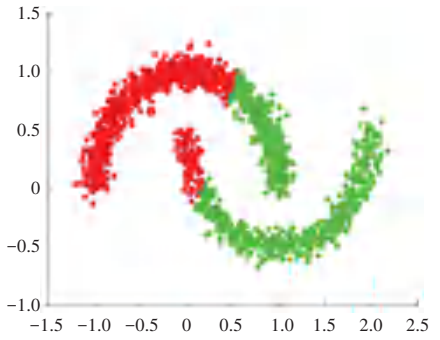
表 7　在 Twomoons 数据集上的 CVI_{XB}，*Accuracy*，*NMI*，*Recall*，*Precision* 和 F_1-value

Tab. 7　CVI_{XB}，*Accuracy*，*NMI*，*Recall*，*Precision* and F_1 value on Twomoons

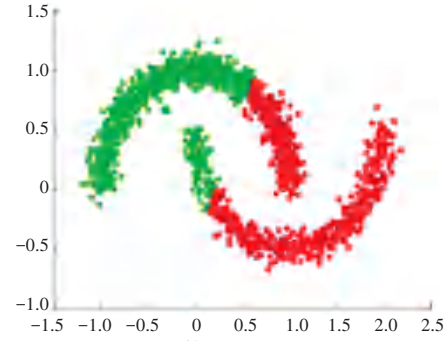
算法	CVI_{XB}	*Accuracy*	*NMI*	*Recall*/%	*Precision*/%	F_1 - value
FCM	0.141 3	0.731 0	0.1761	69.33	78.31	0.735 5
AFCM-SP	**0.095 9**	0.715 9	0.177 9	65.43	80.18	0.720 6
SFCM	0.223 4	0.732 4	0.191 6	66.48	**82.83**	0.737 6
Hybrid FCM	0.125 3	**0.735 0**	**0.199 5**	**70.03**	78.39	**0.739 7**



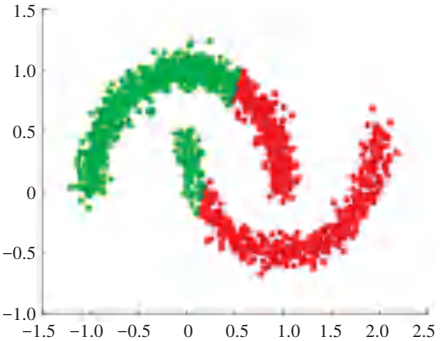
（a）FCM 算法聚类后的分布图



（b）AFCM-SP 算法聚类后的分布图



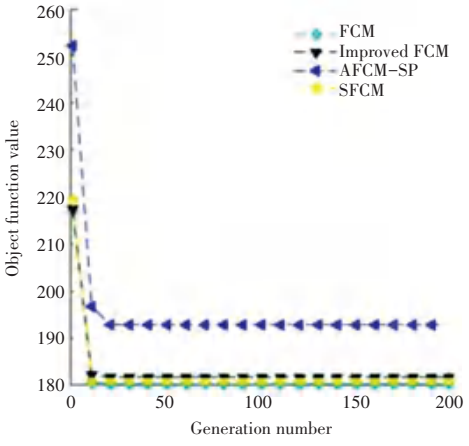
（c）SFCM 算法聚类后的分布图



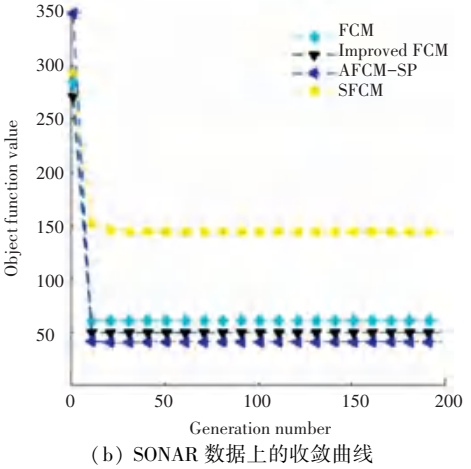
（d）Hybrid FCM 算法聚类后的分布图

图 1　在 Twomoons 数据集上的聚类分布情况

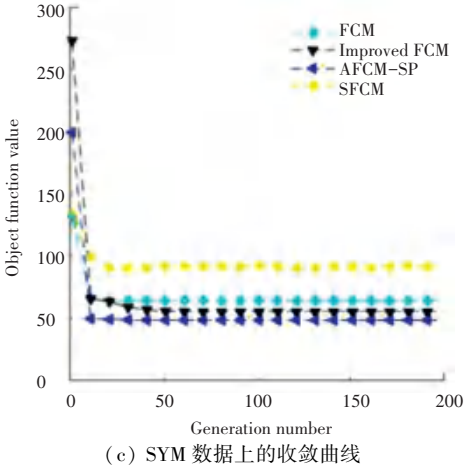
Fig. 1　Clustering distribution on Twomoons dataset



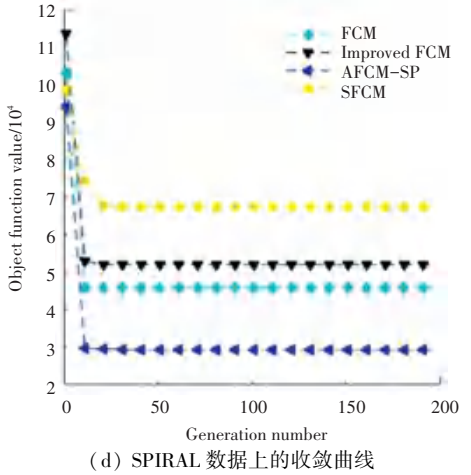
（a）IRIS 数据上的收敛曲线



（b）SONAR 数据上的收敛曲线



（c）SYM 数据上的收敛曲线



（d）SPIRAL 数据上的收敛曲线

图 2　不同数据集上各算法的收敛曲线

Fig. 2　Convergence curves of each algorithm on different datasets

4　结束语

针对传统 FCM 算法依赖于初始聚类中心、对噪声敏感、容易陷入局部最优等缺点，本文提出一种基于余弦相似性的自适应权重的改进 FCM 算法。首先，新算法考虑了样本与聚类中心联合起来对目标函数的影响程度，引入了样本与聚类中心权重和相应的自适应因子，使得原问题解空间的维数更大，最优解的精度提高；其次，提出了一种基于权重的聚类有效性指标作为 PSO 算法的适应度函数去优化模糊因子 m 与自适应因子 p，q；最后，对隶属度函数进行余弦相似性修正，大大增强了算法的鲁棒性。实验结果表明本文改进的算法能有效提高算法的聚类性能。

参考文献

[1] DUNN J C. A fuzzy relative of the ISODATA process and its use in detecting compact well－separated clusters[J]. Journal of Cybernetics, 1973,3(3): 32-57.

[2] 夏邢, 薛涛, 李婷. 基于 Spark 的模糊 C 均值算法改进[J]. 西安工程大学学报, 2019, 33(1):100-105.

[3] 徐晓东,吕干云,鲁涛, 等. 基于智能电表数据与模糊 C 均值算法的台区识别[J]. 南京工程学院学报(自然科学版),2020,18(4):1-7.

[4]高立扬, 牛衍亮, 张小平. 基于模糊 C 均值聚类推理模型的高铁土建工程造价智能估算[J]. 石家庄铁道大学学报(社会科学版), 2020, 14(2):36-43.

[5] XIAO Mansheng, WEN Zhicheng, ZHANG Juwu, et al. An FCM clustering algorithm with improved membership function[J]. Control and Decision, 2015,30(12):2270-2274.

[6] ZHOU Wengang, SUN Ting, ZHU Hai. Image segmentation algorithm based on FCM optimized by adaptive spatial information[J]. Application Research of Computers, 2015, 32(7): 2205-2208.

[7] YANG Xiyang, YU Fusheng, PEDRYCZ W. Typical characteristics-based type-2 fuzzy C-Means algorithm[EB/OL]. [2020]. https://10.1109/TFUZZ.2020.2969907.

[8] HESAM I, AJITH A. Fuzzy C-means and fuzzy swarm for fuzzy clustering problem[J]. Expert Systems with Applications,2011,38(3):1835-1838.

[9] 文传军, 詹永照. 粒子群高斯诱导核模糊 C 均值聚类算法[J]. 科学技术与工程, 2018, 18(8):78-84.

[10] 鲁明, 王彬, 刘东儒,等. 基于蚁群优化模糊 C 均值聚类算法的疲劳驾驶研究[J]. 湖北汽车工业学院学报, 2019, 33(2):23-28.

[11]WU Ziheng, WU Zhongcheng, ZHANG Jun. An improved FCM algorithm with adaptive weights based on SA-PSO[J]. Neural Computing and Applications, 2017,28: 3113-3118.

[12]XIE X L, BENI G. A validity measure for fuzzy clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1991, 13(8):841-847.

[13]刘军梅. 新型混沌粒子群混合优化算法[J].软件导刊,2017,16(2):59-62.

[14]徐超, 单志勇, 徐好好. 具有动态学习能力的分层进化粒子群优化算法[J]. 软件导刊, 2021, 20(1):128-131.

[15]LIU Weibo, WANG Zidong, LIU Xiaohui, et al. A novel particle swarm optimization approach for patient clustering from emergency departments[J]. IEEE Transactions on Evolutionary Computation, 2019,23(4): 632-644.

[16] LI Minxuan. An improved FCM clustering algorithm based on cosine similarity[C]// ACM International Conference Proceeding Series. Hong Kong,China:ACM,2019: 103-109.

[17]LANCICHINETTI A , FORTUNATO S , KERTéSZ J. Detecting the overlapping and hierarchical community structure of complex networks[J]. New Journal of Physics, 2009, 11(3):033015.

(上接第 72 页)

[7] KIM Y, MUN S, YOO S, et al. Precise learn-to-rank fault localization using dynamic and static features of target programs[J]. ACM Transactions on Software Engineering and Methodology (TOSEM), 2019, 28(4): 1-34.

[8] ZOU Daming, LIANG Jingjing, XIONG Yingfei, et al. An empirical study of fault localization families and their combinations[J]. IEEE Transactions on Software Engineering, 2019,47(2): 332-347.

[9] XIE Xiaoyuan, CHEN T Y, KUO F-C, et al. A theoretical analysis of the risk evaluation formulas for spectrum-based fault localization[J]. ACM Transactions on Software Engineering and Methodology (TOSEM), 2013, 22(4): 1-40.

[10] YOO S. Evolving human competitive spectra-based fault localisation techniques[M]//FRASER G, de SOUZA J T. Search based software engineering. SSBSE 2012. Lecture Notes in Computer Science. Berlin/ Heidelberg:Springer, 2012,7515: 244-258.

[11]XUAN Jifeng, MONPERRUS M. Learning to combine multiple ranking metrics for fault localization[C]// 2014 IEEE International Conference on Software Maintenance and Evolution. Victoria, BC, Canada:IEEE, 2014: 191-200.

[12]PENG Hanchuan, LONG Fuhui, DING C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy[J]. IEEE Transactions on pattern analysis and machine intelligence, 2005, 27(8): 1226-1238.

[13] PEDREGOSA F, VAROQUAUX G, GRAMFORT A, et al. Scikit-learn: Machine learning in Python[J]. the Journal of machine Learning research, 2011, 12: 2825-2830.

[14]JUST R, JALALI D, ERNST M D. Defects4J: A database of existing faults to enable controlled testing studies for Java programs[C]// Proceedings of the 2014 International Symposium on Software Testing and Analysis. San Jose, CA, USA : ACM, 2014: 437-440.

[15]JIANG Bo, CHAN W K, TSE T H. On practical adequate test suites for integrated test case prioritization and fault localization[C]// 2011 11th International Conference on Quality Software. Madrid, Spain:IEEE, 2011: 21-30.