Advances in Applied Mathematics 应用数学进展, 2021, 10(4), 953-962
Published Online April 2021 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2021.104104

Hans汉斯

基于PSO-TVAC的中心自适应权的FCM聚类算法

胡建华，尹慧琳

上海理工大学理学院，上海
Email: hjh_2021@usst.edu.cn, 1059102134@qq.com

收稿日期：2021年3月15日；录用日期：2021年4月3日；发布日期：2021年4月20日

摘　要

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、分类时会倾向于多数类等缺点，本文首先提出一种基于PSO-TVAC的中心自适应权的FCM聚类算法(CWAFCM)。新算法将中心权重向量φ和自适应指数q引入目标函数，用以区分每个聚类中心的不同重要性；指数q和模糊因子m由粒子群算法(PSO-TVAC)优化；新提出一种聚类评价标准ACVI作为PSO-TVAC算法的适应度函数以提高聚类准确率。其次，将CWAFCM与过采样技术(SMOTE)相结合以适应于对不平衡数据聚类。通过对六个数据集(四个平衡数据集，两个不平衡数据集)进行仿真实验，结果表明CWFCM算法能够有效地优化聚类效果，且能提高不平衡数据集的聚类准确率。

关键词

模糊c均值算法，自适应权重，过采样技术，粒子群算法

FCM Clustering Algorithm Based on PSO-TVAC Algorithm with Adaptively Weighted Centers

Jianhua Hu, Huilin Yin

College of Science, University of Shanghai for Science and Technology, Shanghai
Email: hjh_2021@usst.edu.cn, 1059102134@qq.com

Received: Mar. 15th, 2021; accepted: Apr. 3rd, 2021; published: Apr. 20th, 2021

Abstract

The traditional FCM algorithm relies on the initial clustering center, is sensitive to noise, is easy to

文章引用: 胡建华, 尹慧琳. 基于 PSO-TVAC 的中心自适应权的 FCM 聚类算法[J]. 应用数学进展, 2021, 10(4): 953-962.
DOI: 10.12677/aam.2021.104104

fall into local optimum, and tends to classify most classes. In this paper, a FCM clustering algorithm based on PSO-TVAC algorithm with adaptively weighted centers is proposed. The new algorithm introduces the weight vector φ of centers and the adaptive exponent q into the objective function to distinguish the different importance of each cluster center. The exponent q and fuzzy factor m are optimized by particle swarm optimization (PSO-TVAC). Secondly, CWAFCM is combined with synthetic minority oversampling technique (SMOTE) to cluster unbalanced data. The results of experiments on six datasets (four balanced datasets and two unbalanced datasets) show that CWAFCM algorithm can effectively optimize the clustering effect and improve the clustering accuracy on unbalanced dataset.

Keywords

Fuzzy c-Means Algorithm (FCM), Adaptive Weight, Oversampling Technology, Particle Swarm Optimization

1. 引言

FCM 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，并能获得较好的结果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] [3] [4]。但 FCM 算法仍然存在依赖初始聚类中心、对噪声敏感度、容易陷入局部最优等缺陷。近些年来，许多改进的 FCM 算法[5] [6] [7]和聚类有效性指标被提出以提高聚类的效果。如 Ayan Seal 等人[8]提出一种基于 Jeffreys 散度为相似度度量的改进 FCM 算法(JDFCM)，大大提高了聚类性能。Niteesh Kumar 等人[9]给出两种不同的距离度量方法来适应噪声环境，提出了 AMFCM 算法和 EMFCM 算法。同时，FCM 聚类算法与其他优化算法的结合也大大提高了算法的性能。例如，文献[10]讨论了一种基于 FCM 和 FPSO 的混合聚类算法(FCMPSO)用来提高聚类效果。因为在噪声和样本分布不均衡的环境下，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的 FCM 算法(AFCM)。

众所周知，聚类中心在聚类过程中起着重要的作用，对聚类结果有着重要的影响。事实上，每个聚类中心都有自己的重要性，应该区别对待。例如，在对中国城市群的研究中，北京、上海这样的中心城市，由于其巨大的经济辐射能力，应该比西安、成都等西部中心城市更受到关注。为此，本文提出一种基于 PSO-TVAC 的中心自适应权的 FCM 聚类算法(CWAFCM)。新算法将中心权重向量 ψ 和自适应指数 q 引入目标函数，用以区分每个聚类中心的不同重要性；指数 q 和模糊因子 m 由粒子群优化算法(PSO-TVAC)优化所确定以降低陷入局部最优的可能性；为恰当刻画类内的紧致性和类间分离程度，新提出一种聚类评价标准 ACVI，并作为 PSO-TVAC 算法的适应度函数以提高聚类准确率。其次，针对传统的聚类算法分类时会倾向于多数类的缺点，本文将新提出的 CWAFCM 算法与过采样技术(SMOTE)相结合以适应于对不平衡数据聚类。通过对六个数据集(四个平衡数据集，两个不平衡数据集)进行仿真实验，结果表明 CWFCM 算法能够有效地优化聚类效果，且能提高不平衡数据集的聚类准确率。

2. 传统的 FCM 算法与 SMOTE 技术

2.1. 模糊 c 均值算法(FCM)

FCM 算法主要通过迭代方式最小化其目标函数以获得样本集的模糊划分，它将靠近聚类中心的样本

点赋予较高的隶属度，而远离聚类中心的样本点赋予较低的隶属度。设 $X=\{x_1,x_2,\cdots,x_n\}$ 是数据集，n 是样本数，c 是已知的类别数。传统的 FCM 算法以欧氏距离作为相似度度量，其目标函数为：

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

其中 x_i 表示第 i 个样本点，v_j 表示第 j 个聚类中心，$\|\cdot\|$ 代表欧氏距离，m 为模糊因子，用来控制聚类结果的模糊程度和函数的凸性，一般取值为 2，μ_{ji} 代表第 i 个样本点属于第 j 个聚类中心的隶属度，满足约束条件

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

令 $U=(\mu_{ji})_{c\times n}$ 表示模糊隶属度矩阵，$V=\{v_1,v_2,\cdots,v_c\}$ 是所有聚类中心的集合。FCM 算法通过迭代更新隶属度矩阵与聚类中心以实现聚类的目的，其迭代公式为

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

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

由式(1)，FCM 算法的目标函数可以改写为

$$J=1\cdot\sum_{j=1}^{c}\mu_{j1}^{m}\left\|x_1-v_j\right\|^2+1\cdot\sum_{j=1}^{c}\mu_{j2}^{m}\left\|x_2-v_j\right\|^2+\cdots+1\cdot\sum_{j=1}^{c}\mu_{jn}^{m}\left\|x_n-v_j\right\|^2 \tag{5}$$

即每个样本点和它的类中心之间的模糊距离系数为 1，这说明传统 FCM 算法中每个样本点对目标函数的贡献是同等重要的，即使该样本点是噪声或离群点，显然这与实际情况不符。考虑到每个样本点的不同重要性，文献[11]提出了一种改进的 AFCM 算法，通过引入了一个权重向量 $\psi=(\psi_1,\psi_2,\cdots,\psi_n)$，来刻画每个样本点的重要性。其目标函数为

$$G_a(\psi,U,V)=\sum_{i=1}^{n}\psi_i^{p}\sum_{j=1}^{c}\mu_{ji}^{m}\left\|x_i-v_j\right\|^2 \tag{6}$$

这里每一个 $\mu_{ji}\in[0,1]$，ψ_i 为样本 x_i 的自适应权重，$\psi_i>0$，$i=1,2,\cdots,n$，满足约束条件：

$$\sum_{j=1}^{c}\mu_{ji}=1,\quad \prod_{i=1}^{n}\psi_i=1 \tag{7}$$

参数 p 表示自适应指数，用来控制自适应权值。AFCM 算法的聚类中心和权重向量的迭代公式为式(8)和式(9)，隶属度更新公式与式(3)相同，

$$\psi_i=\left\{\frac{\left[\prod_{i=1}^{n}\left(\sum_{j=1}^{c}\mu_{ji}^{m}\left\|x_i-v_j\right\|^2\right)\right]^{\frac{1}{n}}}{\sum_{j=1}^{c}\mu_{ji}^{m}\left\|x_i-v_j\right\|^2}\right\}^{\frac{1}{p}} \tag{8}$$

$$v_j = \frac{\sum_{i=1}^{n} \psi_i \mu_{ji}^m x_i}{\sum_{i=1}^{n} \psi_i \mu_{ji}^m} \tag{9}$$

AFCM 算法有效减少了噪声点的干扰，也适应于样本分布不均衡的情况。

2.2. SMOTE 技术

现实生活中，常常会产生不平衡的数据集，如广告点击、信用卡欺诈等信息数据，而经典的聚类算法倾向于将样本划分到多数类，这将导致聚类结果不可靠。对于不平衡数据的处理有两个方向：一是数据层面的，通过改变数据分布使类别更为平衡；二是算法层面的，通过改进聚类算法使模型更看重少数类，如在传统分类算法的基础上对不同类别采取不同的加权值。

合成少类过采样技术(SMOTE)是一种基于数据层面的处理不平衡数据的方法。其目的是针对少数类样本的特征进行分析，同时合成新样本加入少数类样本，从而平衡数据分布。SMOTE 算法的流程如下：

1) 在少类样本中选取样本 x_i，计算 x_i 到每个少类样本的欧氏距离，选取 K 个距离最小的样本作为 K 近邻。

2) 根据多类样本与少类样本之间的不平衡比率确定采样倍率 N，对于少类中的每个样本 x_i，从其 K 近邻中选取若干个样本，不妨设为 x_n。

3) 对于每个随机选择出的 x_n 与原样本 x_i，按照式(10)，合成的新样本 x_{new}。

$$x_{new} = x_i + rand(0,1)(x_i - x_n) \tag{10}$$

3. 基于 PSO-TVAC 的中心自适应权的 FCM 算法

AFCM [11]算法启示我们，在聚类过程中将每个样本区别以待可以减少噪声干扰，提高算法性能。同样，FCM 算法中，聚类中心起着非常重要的作用，它决定了最终的聚类结果，但其结果依赖于初始中心的选取。为减少这种依赖性，本文将强调每个中心的重要性，提出一种基于 PSO-TVAC 的中心自适应权的 FCM 聚类算法(CWAFCM)。

3.1. 中心自适应权的 FCM 模型

由式(1)，FCM 算法的目标函数可以改写为

$$J = 1\cdot\sum_{j=1}^{c}\mu_{j1}^m \left\|x_1 - v_j\right\|^2 + 1\cdot\sum_{j=1}^{c}\mu_{j2}^m \left\|x_2 - v_j\right\|^2 + \cdots + 1\cdot\sum_{j=1}^{c}\mu_{jn}^m \left\|x_n - v_j\right\|^2, \tag{11}$$

即每个类中心和所有样本点之间的模糊距离系数为 1，这说明传统的 FCM 算法中每个中心都被平等看待，这将最终影响聚类有效性。本文将通过中心引入权重的方法来提高算法的鲁棒性和聚类效果。CWAFCM 算法的目标函数为

$$G_c(\varphi, U, V) = \sum_{j=1}^{c}\varphi_j^q \sum_{i=1}^{n}\mu_{ji}^m \left\|x_i - v_j\right\|^2 \tag{12}$$

其中 φ_j 表示第 j 个聚类中心的权重，$\varphi_j, j = 1,2,\cdots,c$ 组成中心自适应权重向量 φ，q 为中心权重的自适应指数。满足 $\sum_{j=1}^{c}\mu_{ji} = 1$，$\prod_{j=1}^{c}\varphi_j = 1$。用拉格朗日插值法得到拉格朗日函数为

$$L(\varphi, U, V) = \left(\sum_{j=1}^{c}\varphi_j^q \sum_{i=1}^{n}\mu_{ji}^m \left\|x_i - v_j\right\|^2\right) + \beta\left(\prod_{j=1}^{c}\varphi_j - 1\right) + \sum_{i=1}^{n}\lambda_i\left(\sum_{j=1}^{c}\mu_{ji} - 1\right)$$

在 $L(\varphi,U,V)$ 中对 U,φ,V 求偏导并令其等于 0 有

$$\frac{\partial L}{\partial \mu_{ji}} = m\varphi_j^q \mu_{ji}^{m-1} \left\| x_i - v_j \right\|^2 + \lambda_i = 0$$

$$\frac{\partial L}{\partial \varphi_j} = q\varphi_j^{q-1} \sum_{i=1}^{n} \mu_{ji}^m \left\| x_i - v_j \right\|^2 + \beta \prod_{l=1,l\neq j}^{c} \varphi_l = 0$$

$$\frac{\partial L}{\partial v_j} = -\sum_{j=1}^{c} \varphi_j^q \mu_{ji}^m \left(x_i - v_j \right) = 0$$

因为有 $\prod_{l=1,l\neq j}^{c} \varphi_l = \frac{1}{\varphi_j}$，代入上式可以得到使式(12)达到最小的充要条件

$$\mu_{ji} = \frac{1}{\sum_{k=1}^{c} \left[\frac{\varphi_j \left\| x_i - v_j \right\|^2}{\varphi_k \left\| x_i - v_k \right\|^2} \right]^{\frac{1}{m-1}}} \tag{13}$$

$$\varphi_j = \left\{ \frac{\left[\prod_{j=1}^{c} \left(\sum_{i=1}^{n} \mu_{ji}^m \left\| x_i - v_j \right\|^2 \right) \right]^{\frac{1}{c}}}{\sum_{i=1}^{n} \mu_{ji}^m \left\| x_i - v_j \right\|^2} \right\}^{\frac{1}{q}} \tag{14}$$

式(13)和式(14)是隶属度和中心权重的更新迭代公式，聚类中心的更新公式则与式(4)一样。比较式(3)和(13)，CWAFCM 模型中的隶属度随权向量 φ 的改变而改变。

3.2. 聚类有效性指标

聚类有效性指标是一种用来度量聚类效果的评价函数。常用的有效性指标有 Xie Beni 指标(CVI_{XB}) [12]，标准互信息(NMI) [13]，聚类准确率(Accuracy)，召回率，F1 值等。考虑到 FCM 聚类过程中不同中心的重要性不同，本文提出一种改进的 Xie-Beni 指标 ACVI。令

$$\mathrm{ACVI} = \min_{j\neq k} \left\| \varphi_j v_j - \varphi_k v_k \right\|^2$$

$$\mathrm{ACOMP} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{c} \mu_{ji}^m \left\| x_i - v_j \right\|^2$$

$$\mathrm{ACVI} = \frac{\mathrm{ACOMP}}{\mathrm{ASPT}} = \frac{\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{c} \mu_{ji}^m \left\| x_i - v_j \right\|^2}{\min_{j\neq k} \left\| \varphi_j v_j - \varphi_k v_k \right\|^2} \tag{15}$$

ASPT 为类间最小加权距离，反映类间分离程度，其值越大越好；ACOMP 为类内平均模糊距离，反映类内紧致程度，其值越小越好。因此作为比值，ACVI 的值越小表示聚类效果越好。

3.3. PSO-TVAC 算法

CWAFCM 算法中，模糊因子 m 和自适应指数 q 是影响聚类结果的关键参数。恰当取值不仅可以降低算法陷入局部最优的可能性，还能提高聚类准确度。带时变加速度的粒子群算法(PSO-TVAC) [14]因其算法简单，搜索速度快，效率高等优点被广泛用于参数优化问题。本文采用 PSO-TVAC 来优化参数 m 和 q，其适应度函数采用本文提出的聚类指标 ACVI。PSO-TVAC 算法的速度和位置更新公式如下：

$$\boldsymbol{V}_i(t+1)=\omega\boldsymbol{V}_i(t)+c_1\boldsymbol{r}_1\left(\boldsymbol{P}_i(t)-\boldsymbol{X}_i(t)\right)+c_2\boldsymbol{r}_2\left(\boldsymbol{P}_g(t)-\boldsymbol{X}_i(t)\right) \tag{16}$$

$$\boldsymbol{X}_i(t+1)=\boldsymbol{X}_i(t)+\boldsymbol{V}_i(t+1) \tag{17}$$

其中$\boldsymbol{V}_i(t)$和$\boldsymbol{X}_i(t)$分别表示搜索空间中第 t 次迭代时的速度$\boldsymbol{V}_i$和位置 $\boldsymbol{X}_i$，ω 为惯性权重，c_1,c_2 为加速度系数，计算公式为

$$\omega=\omega_{\max}-\frac{\omega_{\max}-\omega_{\min}}{\text{iter}_{\max}}*\text{iter} \tag{18}$$

$$c_1=\left(c_{1i}-c_{1f}\right)*\frac{\text{iter}_{\max}-\text{iter}}{\text{iter}_{\max}}+c_{1f} \tag{19}$$

$$c_2=\left(c_{2i}-c_{2f}\right)*\frac{\text{iter}_{\max}-\text{iter}}{\text{iter}_{\max}}+c_{2f} \tag{20}$$

其中 $\omega_{\max}=0.9,\omega_{\min}=0.4$，iter 为迭代次数，$\text{iter}_{\max}$ 为最大迭代次数；c_{1i}，c_{2i}（c_{1f}，c_{2f}）表示 c_1，c_2 初值(终值)，$c_{1i}=c_{2f}=2.5$，$c_{2i}=c_{1f}=0.5$。$\boldsymbol{r}_1,\boldsymbol{r}_2$ 为分量在(0, 1)之间的随机向量，$\boldsymbol{P}_i(t)$表示在 t 时刻的第 i 个粒子的历史最佳位置，$\boldsymbol{P}_g(t)$表示 t 时刻的全局最佳位置。

3.4. CWAFCM 算法的流程

这里我们给出 CWAFCM 算法的流程：首先用 PSO-TVAC 算法优化参数 m, q，采用适应度函数 ACVI；其次以式(12)为目标函数进行聚类。具体步骤为：

输入：数据集和聚类中心数 c；聚类最大迭代次数 Itermax；PSO 算法最大迭代次数 $\text{Iter}_{\max}$。

输出：聚类结果

1) 初始化关于 m, q 的粒子种群 X；初始化隶属度矩阵 $\boldsymbol{U}$，中心权重向量 $\boldsymbol{\varphi}$；计算聚类中心向量 $\boldsymbol{V}$；设置相关参数，收敛阈值 eps；

2) 随机初始化每个粒子的速度 $\boldsymbol{V}_i$ 和位置 $\boldsymbol{X}_i$；

3) 根据式(13)、(14)、(4)更新 $\boldsymbol{U}, \boldsymbol{\varphi}, \boldsymbol{V}$；

4) 用式(15)计算每个粒子的适应度值 ACVI($\boldsymbol{X}_i(t)$)，并得到 $\boldsymbol{P}_i$ 和 $\boldsymbol{P}_g$；

5) 根据公式(18)、(19)和(20)计算 ω,c_1,c_2；用式(16)和(17)更新粒子的速度 $\boldsymbol{V}_i$ 和位置 $\boldsymbol{X}_i$；

6) 当达到最大迭代次数 $\text{Iter}_{\max}$ 或者 $\left|\text{ACVI}\left(\boldsymbol{P}_g(k)\right)-\text{ACVI}\left(\boldsymbol{P}_g(k-1)\right)\right|<\text{eps}$ 时，回到第三步，直到达到 $\text{Iter}_{\max}$；

7) 得到全局最优粒子 $\boldsymbol{P}_g=(m,q)$；

8) 初始化隶属度函数 $\boldsymbol{U}$ 和中心权重向量 $\boldsymbol{\varphi}$，通过式(4)计算聚类中心 $\boldsymbol{V}$；

9) 通过式(4)、(13)、(14)更新聚类中心，隶属度和中心权重；

10) 当整个算法收敛时则得到最终聚类中心和每个粒子的隶属度。

4. 仿真实验与结果分析

4.1. 数据集与对比算法

为了验证新提出的算法的性能，本文选用了四个平衡数据和两个不平衡的数据集进行实验，见表 1。对于不平衡数据，事先用 SMOTE 技术改变数据分布使类别平衡。为说明改进的算法有较好的聚类能力，传统的 FCM 算法[15]，改进的 AFCM 算法[11]，最近的 JDFCM 算法[8]用来作对比实验。

Table 1. Distribution characteristics of six datasets
表 1. 六个数据集的分布特征

数据集	样本数	特征数	类别数目	数据类型
IRIS	150	4	3	平衡
SONAR	208 $\xi^{r(k)}=2$	60	2	平衡
GLASS	214	9	6	平衡
WEKA	310	6	3	平衡
Twomoons	1502	2	2	不平衡
Spiral	567	2	2	不平衡

4.2. 实验参数的设置

文本提出改进的 FCM 算法的仿真过程分为两部分，首先是利用 PSO-TVAC 算法对模糊因子 m、自适应参数 q 进行优化，得到最优的全局最优粒子 $\boldsymbol{P}_g=(m,q)$，然后基于最优的 m, q 值通过 CAFCM 算法得到最终聚类结果。在实验过程中，每个样本隶属度初始值是[0, 1]之间的随机数，中心权重向量 φ 初始值为全 1 向量；聚类过程中的最大迭代次数 $\text{Iter}_{\max}$ 为 200，收敛阈值 eps 为 10^{-5}。在 PSO 算法的过程中，最大迭代次数 $\text{Iter}_{\max}$ 设置为 20，种群大小为 20，粒子在二维搜索空间中移动。由于算法的结果与随机产生的初始值有关，本文中的所有实验都是独立重复 30 次，以消除随机性的影响，然后给出平均值。在每个数据集上，通过 PSO-TVAC 算法优化得到的 m, q 的值列在表 2 中。

Table 2. m, q values by PSO_TVAC algorithm
表 2. 由 PSO-TVAC 算法确定的 m, q 值

数据集	m	q
IRIS	2.49	1.65
SONAR	1.98	2.41
GLASS	2.01	2.34
WEKA	2.04	2.31
Twomoons	1.03	2.50
Spiral	1.34	2.76

4.3. 实验结果分析

为了实验对比的公平性，本文选取常见聚类有效性指标：Xie Beni 指标(CVI_{XB})、聚类准确率(Accuracy)、标准互信息(NMI)。CVI_{XB} 值越小、Accuracy 和 NMI 越大说明聚类效果越好。而对不平衡数据，增加精确率，召回率和 F1 值三个指标以说明新算法可以降低对多数类的倾向性。实验结果由表 3~8 和图 1 给出。在数据集 IRIS 上，新算法 CWAFCM 在指标 CVI_{XB} 和 NMI 都优于另外几种算法，但 Accuracy 上略差于 JDFCM 而与 AFCM 相同；在 SONAR 上，CWAFCM 在 Accuracy 和 NMI 指标上都明显优于另外三种算法，CVI_{XB} 值在四种算法中排第二；在 GLASS 中，CWAFCM 的 Accuracy 和 NMI 都排名第一；在 WEKA 中，CWAFCM 有较为优异的 Accuracy 值，NMI 略逊于 AFCM 算法。对于不平衡数据集，在 Twomoons 上，CWAFCM 仅有 CVI_{XB} 和 Precision 略逊于 AFCM。在 Spiral 上，CWAFCM 仅有 Precision 低于 AFCM。为进一步说明实验结果的有效性，图 1 给出了不同的算法在四个不同数据集上的收敛曲线，

曲线表明每种算法都有较快的收敛速度。综上所述，新算法 CWAFCM 在六个数据上相对于三种对比算法有较好的聚类效果。

Table 3. CVI_{XB}, Accuracy, NMI on IRIS
表 3. 在 IRIS 上的 CVI_{XB} Accuracy, NMI

算法	CVI_{XB}	Accuracy	NMI
FCM	0.1369	0.8933	0.7465
AFCM	0.1631	0.9067	0.7441
JDFCM	0.6244	**0.9267**	0.4445
CWAFCM	**0.0241**	0.9067	**0.7487**

Table 4. CVI_{XB}, Accuracy, NMI on SONAR
表 4. 在 SONAR 上的 CVI_{XB} Accuracy, NMI

算法	CVI_{XB}	Accuracy	NMI
FCM	2.1876	0.5529	0.0088
AFCM	3.0530	0.5529	0.0090
JDFCM	**0.7018**	0.5529	0.0068
CWAFCM	2.0041	**0.5577**	**0.0105**

Table 5. CVI_{XB}, Accuracy, NMI on GLASS
表 5. 在 GLASS 上的 CVI_{XB} Accuracy, NMI

算法	CVI_{XB}	Accuracy	NMI
FCM	2.3578	0.4910	0.3593
AFCM	4.2614	0.4393	0.3277
JDFCM	**0.3852**	0.5187	0.3393
CWAFCM	2.3933	**0.5514**	**0.3937**

Table 6. CVI_{XB}, Accuracy, NMI on WEKA
表 6. 在 WEKA 上的 CVI_{XB} Accuracy, NMI

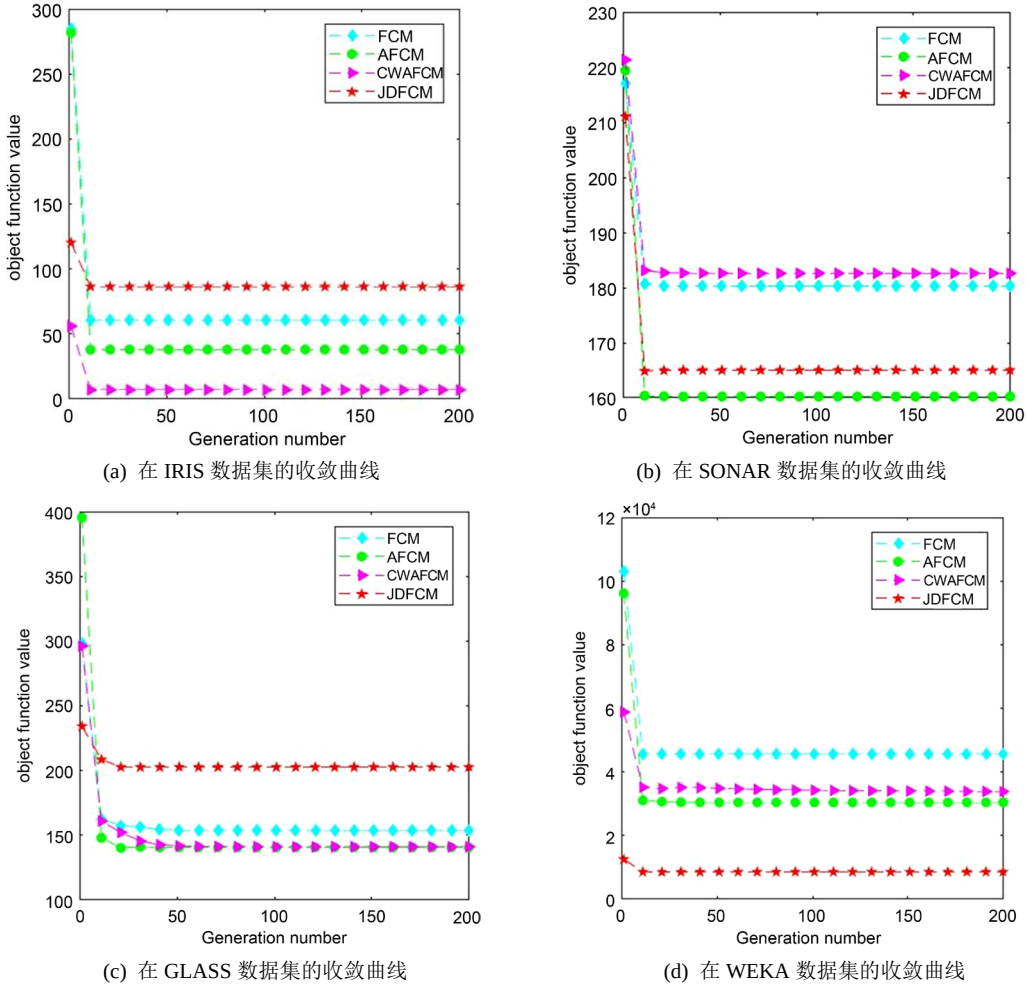
算法	CVI_{XB}	Accuracy	NMI
FCM	**0.3154**	0.5452	0.4172
AFCM	0.4708	0.5548	**0.4224**
JDFCM	0.8072	0.6161	0.1658
CWAFCM	0.4751	**0.6742**	0.4216

Table 7. CVI_{XB}, Accuracy, NMI, Recall, Precision and F1 value on Twomoons
表 7. 在 Twomoons 数据集上的 CVI_{XB}, Accuracy, NMI, Recall, Precision 和 F1-value

算法	CVI_{XB}	Accuracy	NMI	Recall	Precision	F1-value
FCM	**0.1423**	0.7310	0.1761	69.33%	78.31%	0.7355
AFCM	0.0968	0.7117	0.1713	64.30%	**81.10%**	0.7173
JDFCM	0.1566	0.6818	0.0152	67.74%	75.70%	0.7150
CWAFCM	0.1765	**0.7350**	**0.1812**	**69.83%**	78.61%	**0.7396**

Table 8. CVI_{XB}, Accuracy, NMI, Recall, Precision and F1 value on Spiral
表 8. 在 Spiral 数据集上的 CVI_{XB}, Accuracy, NMI, Recall, Precision 和 F1-value

算法	CVI_{XB}	Accuracy	NMI	Recall	Precision	F1-value
FCM	0.2808	0.5839	0.2441	62.16%	57.54%	0.5975
AFCM	0.3012	0.5882	**0.2788**	62.46%	56.11%	0.5911
JDFCM	0.3654	0.5898	0.1575	62.33%	56.84%	0.5944
CWAFCM	**0.2244**	**0.6024**	0.2768	**63.64%**	**58.56%**	**0.6099**



(a) 在 IRIS 数据集的收敛曲线　(b) 在 SONAR 数据集的收敛曲线

(c) 在 GLASS 数据集的收敛曲线　(d) 在 WEKA 数据集的收敛曲线

Figure 1. Convergence curves of each algorithm on different datasets
图 1. 不同数据集下各算法的收敛曲线

5. 结论

传统的 FCM 算法表现出依赖于初始聚类中心、对噪声敏感、容易陷入局部最优、分类时会倾向于多数类等不足。考虑到每个聚类中心在聚类过程中不同的重要性，本文提出一种基于 PSO-TVAC 的中心自适应权的 FCM 聚类算法(CWAFCM)。每个聚类中心的不同重要性用权重向量 φ 来刻画，自适应指数 m, q 来控制目标函数的凸性和聚类的模糊性，作为关键参数它们由 PSO-TVAC 算法所确定；为提高聚类准确度，一种新的聚类评价标准 ACVI 作为 PSO-TVAC 算法的适应度函数。对于不平衡数据，本文将 CWAFCM

与过采样技术(SMOTE)相结合，以实现对不平衡数据的聚类。通过在六个数据集上(四个平衡数据集，两个不平衡数据集)和三个对比算法的对比实验，结果表明 CWAFCM 算法能够有效地优化聚类效果，且能提高不平衡数据集的聚类准确率。这说明新算法能有效地降低对初始中心的依赖和噪声的干扰，减少陷入局部最优的可能。

基金项目

本项目由国家自然科学基金项目：61873169 资助。

参考文献

[1] Dunn, J.C. (1973) A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics*, **3**, 32-57. https://doi.org/10.1080/01969727308546046

[2] 赵战民, 朱占龙, 王军芬. 改进的基于灰度级的模糊 C 均值图像分割算法[J]. 液晶与显示, 2020, 35(5): 499-507.

[3] 冯国政, 徐金东, 范宝德, 赵甜雨, 朱萌, 孙潇. 基于半监督模糊 c 均值算法的遥感影像分类[J]. 计算机应用, 2019, 39(11): 3227-3232.

[4] Huang, H., *et al*. (2019) Brain Image Segmentation Based on FCM Clustering Algorithm and Rough Set. *IEEE Access*, **7**, 12386-12396. https://doi.org/10.1109/ACCESS.2019.2893063

[5] Jun, Y., *et al*. (2017) An Adaptive Clustering Segmentation Algorithm Based on FCM. *Turkish Journal of Electrical Engineering and Computer Sciences*, **25**, 4533-4544. https://doi.org/10.3906/elk-1607-103

[6] Kannan, S.R., Devi, R., Ramathilagam, S. and Takezawa, K. (2013) Effective FCM Noise Clustering Algorithms in Medical Images. *Computers in Biology and Medicine*, **43**, 73-83. https://doi.org/10.1016/j.compbiomed.2012.10.002

[7] Qamar, U. (2014) A Dissimilarity Measure Based Fuzzy c-Means (FCM) Clustering Algorithm. *Journal of Intelligent and Fuzzy Systems*, **26**, 229-238. https://doi.org/10.3233/IFS-120730

[8] Seal, A., Karlekar, A., Krejcar, O., *et al*. (2020) Fuzzy c-Means Clustering Using Jeffreys-Divergence Based Similarity Measure. *Applied Soft Computing*, **88**, Article ID: 106016. https://doi.org/10.1016/j.asoc.2019.106016

[9] Kumar, N., Kumar, H. and Sharma, K. (2020) Extension of FCM by Introducing New Distance Metric. *SN Applied Sciences*, **2**, 714. https://doi.org/10.1007/s42452-020-2417-9

[10] Izakian, H. and Abraham, A. (2011) Fuzzy C-Means and Fuzzy Swarm for Fuzzy Clustering Problem. *Expert Systems with Applications*, **38**, 1835-1838. https://doi.org/10.1016/j.eswa.2010.07.112

[11] Wu, Z.H., Wu, Z.C. and Zhang, J. (2017) An Improved FCM Algorithm with Adaptive Weights Based on SA-PSO. *Neural Computing and Applications*, **28**, 3113-3118. https://doi.org/10.1007/s00521-016-2786-6

[12] Xie, X.L. and Beni, G. (1991) A Validity Measure for Fuzzy Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **13**, 841-847. https://doi.org/10.1109/34.85677

[13] Lancichinetti, A., Fortunato, S. and Kertesz, J. (2009) Detecting the Overlapping and Hierarchical Community Structure in Complex Networks. *New Journal of Physics*, **11**, Article ID: 033015. https://doi.org/10.1088/1367-2630/11/3/033015

[14] Ratnaweera, A., Halgamuge, S.K. and Watson, H.C. (2004) Self-Organizing Hierarchical Particle Swarm Optimizer with Time-Varying Acceleration Coefficients. *IEEE Transactions on Evolutionary Computation*, **8**, 240-255. https://doi.org/10.1109/TEVC.2004.826071

[15] Bezdek, J.C. (1981) Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York. https://doi.org/10.1007/978-1-4757-0450-1