单点逼近型加权模糊 C 均值算法的遥感图像聚类应用

韩　敏　　范剑超
（大连理工大学电子与信息工程学院，大连　116023）

摘　要　针对模糊 C 均值算法对数据分布状态和初始聚类中心过于依赖的问题，利用已知样本信息，提出了一种改进的单点逼近型加权模糊 C 均值算法。该算法首先通过对原始数据进行概率统计和加入样本属性权值来调整数据为均匀分布；然后采用先验样本单点逼近的方法来消除先验样本选取的影响，从而不仅得到了合适的初始聚类中心，而且有效地加快了算法的收敛速度和提高了聚类的精度；最后将改进后算法与遥感数据特点相结合，构成了完整的遥感图像地物聚类算法。通过 UCI 数据集和扎龙湿地遥感数据的试验结果的比较证明，该改进方法是真实有效的。

关键词　聚类分析　模糊 C 均值　初始聚类中心　属性权值
中图法分类号：TP391.41　　**文献标识码**：A　　**文章编号**：1006-8961(2009)11-2333-08

A Single-point Approximation Weighted Fuzzy C-means Clustering Method for Classifying Remote Sensing Images

HAN Min, FAN Jian-chao
(*School of Electronic and Information Engineering*, *Dalian University of Technology*, *Dalian* 116023)

Abstract　Focusing on the fuzzy C-means algorithm's problem that the cluster quality is greatly affected by the data distribution and the stochastic initializing the centrals of cluster, a single-point approximation weighted fuzzy C-means algorithm is proposed by using the part of prior samples information. After the probability statistics of original data is conducted, the weights of data attribute are designed to adjust to the uniform distribution, and then are added in the process of cyclic iteration. What's more, in order to significantly improve the convergence speed and the cluster precision, the proper initial cluster centers are chosen by the single adjustment algorithm, which can also overcome the selection influence of prior samples. In addition, combined with the characteristics of remote sensing data, the modified algorithm is updated for remote sensing image cluster. With the comparison experiment of the UCI data sets and the Zhalong wetland remote sensing data, the real validity of proposed algorithm is proved.

Keywords　cluster analysis, fuzzy C-means, initial centre of cluster, attribute of the weight

1　引　言

由于无监督聚类方法不需要训练样本，仅凭地物地磁辐射强弱在遥感图像上所反映的光谱信息，即可将数据按其自然分布特性进行聚类，并能自动判别地物类别，从而可克服人工解译所带来的主观性因素影响[1]。经常使用的无监督聚类方法有：K-

基金项目：国家科技支撑计划项目(2006BAB14B05)；国家重点基础研究发展计划(973)项目(2006CB403405)；国家自然科学基金项目(60674073)；国家高技术研究发展计划(863)项目(2007AA04Z158)
收稿日期：2008-06-26；**改回日期**：2008-10-16

第一作者简介：韩　敏(1959～　)，女。1999 年于日本九州国立大学获博士学位。现为大连理工大学电子与信息工程学院教授，博士研究生导师。研究方向为神经网络、专家系统、3S 系统及混沌序列分析。E-mail：minhan@dlut.edu.cn

means, Gauss expectation maximum[2] 和模糊 C 均值(fuzzy C-means, FCM)[3]等。地球表层信息的复杂性和开放性决定了对遥感信息的分析具有不确定性和模糊性。FCM 算法因为具有全面考虑影像单元每个像素属于各类别的隶属度的优势,所以能够更好地区分不同地物的类别,并成为了当前遥感图像研究和应用的热点[4]。

FCM 算法在模糊聚类中,对初始化参数十分敏感[5],而且对于大样本数据集遥感图像的处理过程来说,整个算法的收敛速度和地物聚类结果的有效性,很大程度上取决于初始聚类中心的选择[6]。图 1(a)为扎龙保护区湿地原始遥感图像,根据实地采样可知,聚类地物种类 $C=9$。对不同的初始聚类中心的 FCM 聚类结果分别如图 1(b)和图 1(c)所示。通过目测可以明显发现,图 1(b)较好地将遥感图像上的地物分成了 9 类,且地物边界清晰。而图 1(c)则将原始遥感图像粗糙地分为两类,不但不能正确表示聚类结果,而且陷入了局部最小。

(a) 原始遥感图 (b) 正常聚类图 (c) 陷入局部最小图

图 1 初始聚类中心对聚类的影响

Fig. 1 The influence of initial centers for clustering

目前,国内外对 FCM 算法初始聚类中心的研究已经取得了很好的研究成果:Yager 等人利用山峰聚类和减法聚类的方法求取聚类中心[7];张文君等人提出了一种利用正态分布标准差的方法——UFCM(uniform fuzzy C-means)[8],其实质上是初始聚类中心均匀分布的方法;Kim 和 Lee 等人利用摄影手册中出现频率高的点去预测合适的聚类中心(colorchecker fuzzy C-means, CFCM)[9],但是其没有考虑到每一类样本的实际分布情况。

此外,FCM 算法主要是对球状分布的数据具有较好的聚类效果[10],所以 Richard 等人开始研究 FCM 算法中各数据属性权重的影响,并通过调整数据的分布,以达到更高的聚类精度[11],但是对于具体的权值设定和调整过程仍没有较好的方法。

本文为了更好地解决初始聚类中心和数据属性权值设定的问题,充分结合遥感图像通常具有已知部分地面信息的实际特点,提出了一种单点逼近型加权模糊 C 均值算法(single-point approximation weighted fuzzy C-means, SWFCM)。该算法首先以这部分先验知识指导算法的收敛方向,并通过概率统计进行权值设定,然后通过调整数据分布状态来加大数据属性之间的区分程度,从而提高聚类精度。此外,SWFCM 算法通过单点逼近方法,能以较低的计算复杂度选取较好的初始聚类中心来加快算法的收敛速度。

2 单点逼近型加权改进算法

在遥感应用领域,每幅遥感图像都会有部分地物类别标记的数据,而基本 FCM 算法属于无监督聚类算法,无法利用这部分有效信息。本文充分利用较少的真实地面信息来进行模型估计,首先确定了数据的属性权值和初始聚类中心,由于克服了算法初始化和原始数据分布对算法的影响,从而提高了算法聚类的精度和速度。

2.1 聚类问题描述

定义聚类分析的图像像素的数目为 N,设图像像素的集合为 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}^{\mathrm{T}}$,如果把图像全部像素分为 C 个类别,每个类别的聚类中心集合为 $\boldsymbol{V}=\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_C\}^{\mathrm{T}}$,并用 $u_{i,k}$ 表示像素点 $\boldsymbol{x}_k$ 隶属于第 i 个聚类中心的隶属度,则隶属度矩阵为

$$\boldsymbol{U}=[u_{i,k}]_{C\times N} \tag{1}$$

加入模糊加权因子 m,则目标函数为

$$J_m(\boldsymbol{U},\boldsymbol{V})=\sum_{k=1}^{N}\sum_{i=1}^{C}u_{i,k}^{m}(\boldsymbol{x}_k-\boldsymbol{v}_i)^{\mathrm{T}}(\boldsymbol{x}_k-\boldsymbol{v}_i) \tag{2}$$

其中,$(\boldsymbol{x}_k - \boldsymbol{v}_i)^{\mathrm{T}}(\boldsymbol{x}_k - \boldsymbol{v}_i)$ 为 Euclidean 距离。SWFCM 算法的物理含义是使求取 $\boldsymbol{X}$ 同一类内的距离之和最小,而不同类间的距离之差最大的迭代循环过程,即寻找 $J_m(\boldsymbol{U},\boldsymbol{V})$ 的最小值的过程。通常情况下,模糊加权因子 m 的最佳范围[12]是[1.5,2.5],一般令 $m=2$。

矩阵 $\boldsymbol{U}$ 中每一列的元素表示每个样本点隶属于各种地物的隶属度,其满足如下 3 个约束条件:

$$\sum_{k=1}^{N} u_{i,k} > 0; \quad \sum_{i=1}^{C} u_{i,k} = 1 \qquad 0 \leqslant u_{i,k} \leqslant 1 \tag{3}$$

其中,$i=1,2,\cdots,C$;$k=1,2,\cdots,N$,上述 3 个表达式表明,隶属度 $u_{i,k}$ 可取 0 到 1 之间的任意实数。这样,一个像素点可以同时隶属于不同的类别,但其隶属于各类别的隶属度总和等于 1。

2.2　数据属性权值设定

虽然 FCM 算法对于每一个数据属性都赋予相同的权值,但是在进行遥感图像地物聚类时,其对地形数据属性的权值调整十分敏感,然而通过权值调整可以使算法能获得适用特殊情况下的聚类结果[13],具有针对性。因此本文利用较少的地面真实信息进行概率统计,首先获取每个样本数据的属性值的均值 μ 和标准差 σ,进而通过设定数据属性的权值来调整数据的分布状态,使其适应 FCM 算法对球形分布数据具有较高聚类精度的特点,以提高整体的聚类精度。

可根据第 i 个数据属性得到的均值来建立向量 $\boldsymbol{\mu}_i=(\mu_{1,i},\mu_{2,i},\cdots,\mu_{C,i})^{\mathrm{T}}$。首先将均值向量归一化到[0,1]上,以避免发生不同数据属性在不同数量级上的问题,然后从大到小排序,得到新的向量 $\tilde{\boldsymbol{\mu}}_i=(\tilde{\mu}_{1,i},\tilde{\mu}_{2,i},\cdots,\tilde{\mu}_{C,i})^{\mathrm{T}}$。

定义 1　对于给定的类别数 C 和数据属性的维数 P,其相邻数据属性之间的差别为

$$\delta_{j,i} = \left\| \tilde{\mu}_{j+1,i} - \tilde{\mu}_{j,i} \right\|_2 \tag{4}$$

其中,$\delta_{j,i}$ 表示相邻两个排序之后的均值之差,$i=1,2,\cdots,P$,其代表了第 j 类和第 $j+1$ 类的第 i 个数据属性的差别,且 $\sum_{j=1}^{C-1}\delta_{j,i}=1$。

定义 2　对于给定的类别数 C 和数据属性的维数 P 的第 i 个数据属性的类别之间区分度为

$$D(i) = \prod_{j=1}^{C-1} \delta_{j,i}^{t} \tag{5}$$

式中,$t\in[0,1]$ 是非线性的增大差值的指数,因为均值全部归一化之后的差别特别小,所以可利用 t 配合来增加每个数据属性之间的区分度。对式(5)进行变形,即可得到下式:

$$\frac{1}{C-1} = \frac{\sum_{j=1}^{C-1}\delta_{j,i}}{C-1} \geqslant C-1\sqrt{\prod_{j=1}^{C-1}\delta_{j,i}} \tag{6}$$

式中,当且仅当 $\delta_{j,i}, j=1,2,\cdots,C-1$ 都相等时,乘积取最大值,即向量 $\boldsymbol{\mu}_i$ 的元素在[0,1]区间内均匀分布,此时,第 i 个数据属性的每个类别的均值区分度越好。

定义 3　对于给定的类别数 C 和数据属性的维数 P 的第 i 个数据属性与其他数据属性之间的重叠程度为

$$O(i) = \left(\frac{P\sum_{l=1}^{C}\sigma_{l,i}}{\sum_{l=1}^{C}\sum_{q=1}^{P}\sigma_{l,q}} \right)^{-1} \tag{7}$$

基于式(5)和式(7),即可得到以下每个数据属性的权值设定准则:

$$\begin{aligned} w_{i,i} &= \alpha O(i) + (1-\alpha)D(i) \\ &= \alpha\left(\frac{P\sum_{l=1}^{C}\sigma_{l,i}}{\sum_{l=1}^{C}\sum_{q=1}^{P}\sigma_{l,q}} \right)^{-1} + (1-\alpha)\prod_{j=1}^{C-1}\delta_{j,i}^{t} \end{aligned} \tag{8}$$

其中,$i=1,2,\cdots,P$,α 为重叠程度和区分度两个指标对权重影响的调整参数。由 $w_{i,i}$ 即构成了以下每个数据属性的权值矩阵

$$\boldsymbol{W} = \begin{pmatrix} w_{1,1} & 0 & \cdots & 0 \\ 0 & w_{2,2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & w_{p,p} \end{pmatrix} \tag{9}$$

为了简化计算,权值矩阵 $\boldsymbol{W}$ 为对角线矩阵,对角线上的元素表示给每个数据属性所加的权重,其他元素为 0,这就排除了数据属性之间的相关性影响。

从式(8)可以得出,若每个样本属性的均值区分度越高,则所赋予的权值越大。而各类样本的均值差别越大,则越容易通过这个属性对每个类别进行聚类。所以,加大权值相当于扩大了该属性对聚类结果的作用,由于隶属度几乎由权值较大的样本数据属性所决定,从而避免了一些各类别之间相近的属性的影响;此外,标准差越小,其所赋予的权值

越高。标准差较小说明了样本的分布很集中,由于其围绕一个较小的波动范围呈球形分布,从而更有利于模糊聚类算法对数据样本集进行聚类。

2.3 初始聚类中心的单点逼近

如果只通过先验样本得到初始聚类中心,那么最终的聚类结果必然会很大程度上依赖于先验样本选取的好坏。因此,必须对前期得到的先验样本均值进行快速调整,以便得到更接近于实际类别中心的初始聚类中心。

可将 2.2 节得到的训练样本的均值向量记作 $\boldsymbol{M}=(\boldsymbol{m}_1,\boldsymbol{m}_2,\cdots,\boldsymbol{m}_C)^{\mathrm{T}}$,其中 C 为聚类的类别数,$\boldsymbol{m}_j=(m_{j,1},m_{j,2},\cdots,m_{j,P})^{\mathrm{T}}$,$P$ 为每类样本所具有的数据属性维数,在 TM 遥感图像中一般提取的是 RGB 值等 3 个属性。

定义 4 $\eta(\boldsymbol{x}_k,\boldsymbol{m}_j)$ 为第 k 个样本点 $\boldsymbol{x}_k=\{x_1^{(k)},x_2^{(k)},\cdots,x_P^{(k)}\}$ 与训练样本均值 $\boldsymbol{m}_j=(m_{j,1},m_{j,2},\cdots,m_{j,P})^{\mathrm{T}}$ 的差别($j=1,2,\cdots,C$)

$$\eta(\boldsymbol{x}_k,\boldsymbol{m}_j)=\|\boldsymbol{x}_k-\boldsymbol{m}_j\|_2 \tag{10}$$

式中,$\eta(\boldsymbol{x}_k,\boldsymbol{m}_j)$ 越大,表示第 k 个样本点和第 j 类的训练样本中心的差异越大,因而不属于第 j 类;反之,表示样本点 $\boldsymbol{x}_k$ 和第 j 类的训练样本中心非常接近;然后即可利用下式对样本进行粗聚类:

$$Z_j=\min_{1\leqslant i\leqslant C}\{\eta(\boldsymbol{x}_k,\boldsymbol{m}_j)\}$$
$$k=1,2,\cdots,N;j=1,2,\cdots,C \tag{11}$$

其中,j 为最小距离所对应的类别,集合 $\boldsymbol{Z}$ 用于存放粗聚类之后的结果。首先利用式(12)计算每个类别的平均值,在得到了粗略的初始聚类中心 $\boldsymbol{v}_i^{(0)}$ 之后,即可用式(13)和式(14)进行单点调整。

$$\boldsymbol{v}_i^{(0)}=\frac{1}{n_j}\left(\sum_{j=1}^{n_j}\boldsymbol{x}_j\right)\qquad \boldsymbol{x}_j\in\boldsymbol{Z}_i \tag{12}$$

$$\rho_{i,i}=\frac{n_i}{n_i-1}[\boldsymbol{x}_k^{(i)}-\boldsymbol{v}_i^{(0)}]^{\mathrm{T}}\boldsymbol{W}[\boldsymbol{x}_k^{(i)}-\boldsymbol{v}_i^{(0)}] \tag{13}$$

$$\rho_{i,j}=\frac{n_j}{n_j+1}[\boldsymbol{x}_k^{(i)}-\boldsymbol{v}_i^{(0)}]^{\mathrm{T}}\boldsymbol{W}[\boldsymbol{x}_k^{(i)}-\boldsymbol{v}_i^{(0)}] \tag{14}$$

其中,$i=1,2,\cdots,C$ 且 $j=1,\cdots,C,i\neq j$,$\boldsymbol{x}_k^{(i)}$ 表示第 i 类中的第 k 个样本点,n_i 和 n_j 分别代表集合 $\boldsymbol{Z}$ 中第 i 类和第 j 类中的总的样本数目。$\rho_{i,i}$ 表示此样本到第 1 次聚类之后所属类别聚类中心的距离,$\rho_{i,j}$ 表示此样本点到其他类别聚类中心的距离。令 $\rho_{i,l}$ 为 $\rho_{i,j}$ 中的最小值,如式(15)所示,如果 $\rho_{i,l}\neq\rho_{i,i}$,则样本点 $\boldsymbol{x}_k^{(i)}$ 需要从第 i 类调整到第 j 类中,j 代表与最小距离所对应的类别。可利用式(16)和式(17)进行调整:

$$\rho_{i,l}=\min\{\rho_{i,j}\}\qquad j=1,\cdots,C \tag{15}$$

$$\boldsymbol{v}_i^{(0)}=\boldsymbol{v}_i^{(0)}+\frac{1}{n_i-1}[\boldsymbol{v}_i^{(0)}-\boldsymbol{x}_k^{(i)}] \tag{16}$$

$$\boldsymbol{v}_j^{(0)}=\boldsymbol{v}_j^{(0)}-\frac{1}{n_j+1}[\boldsymbol{v}_j^{(0)}-\boldsymbol{x}_k^{(i)}] \tag{17}$$

当遍历完所有的样本点,则初始聚类中心集合 $\boldsymbol{V}^{(0)}$ 停止调整。至此,就得到了较好的聚类中心,可将其代入整个 SWFCM 算法中,作为整个迭代循环的初始聚类中心。

进行单点调整的收敛性能证明如下,参数 I 代表整个算法的循环次数:

$$J_i^{(I)}=\sum_{k=1}^{n_i}D_{k,i}^{(I)}$$

其中,$D_{k,i}^{(I)}=[\boldsymbol{x}_k-\boldsymbol{V}_i^{(I)}]^{\mathrm{T}}\boldsymbol{W}[\boldsymbol{x}_k-\boldsymbol{V}_i^{(I)}]$

$$J_i^{(I+1)}=\sum_{k=1}^{n_{i-1}}[\boldsymbol{x}_k-\boldsymbol{v}_i^{(I+1)}]^{\mathrm{T}}\boldsymbol{W}[\boldsymbol{x}_k-\boldsymbol{v}_i^{(I+1)}]$$

$$=\sum_{k=1}^{n_i-1}\sum_{j=1}^{P}\left[\boldsymbol{x}_{k,j}-\boldsymbol{v}_{i,j}^{(I)}-\frac{1}{n_i-1}(\boldsymbol{v}_{i,j}^{(I)}-\boldsymbol{x}_{k,j}^{(i)})\right]^2w_{j,j}^2$$

$$=\sum_{k=1}^{n_i-1}\sum_{j=1}^{P}(\boldsymbol{x}_{k,j}-\boldsymbol{v}_{i,j}^{(I)})^2w_{j,j}^2-$$
$$\frac{1}{n_i-1}\sum_{j=1}^{P}(\boldsymbol{x}_{k,j}^{(i)}-\boldsymbol{v}_{i,j}^{(I)})^2w_{j,j}^2$$

$$=\sum_{k=1}^{n_i}D_{k,i}^{(I)}-\frac{n_i}{n_i-1}[\boldsymbol{x}_k^{(i)}-\boldsymbol{v}_i^{(I)}]^{\mathrm{T}}\boldsymbol{W}[\boldsymbol{x}_k^{(i)}-\boldsymbol{v}_i^{(I)})]$$

$$=J_i^{(I)}-\rho_{i,i}$$

同理,

$$J_j^{(I+1)}=J_j^{(I)}+\rho_{i,j}$$
$$J_C^{(I+1)}=J_C^{(I)}-(\rho_{i,i}-\rho_{i,j})$$

所以当 $\rho_{i,i}>\rho_{i,j}$ 时,则进行循环计算,以使目标函数可以不断减小。

3 基于 SWFCM 的遥感聚类算法实现

根据 SWFCM 改进算法,即可得到了数据属性的权值矩阵 $\boldsymbol{W}$ 和较接近于实际类别中心的初始聚类中心集合 $\boldsymbol{V}^{(0)}=\{\boldsymbol{v}_1^{(0)},\boldsymbol{v}_2^{(0)},\cdots,\boldsymbol{v}_C^{(0)}\}^{\mathrm{T}}$,当将其代入到 SWFCM 算法的整体循环迭代过程中,则基于 SWFCM 的遥感聚类算法步骤如下:

(1)初始化参数。设定初始参数类别数目 C,模糊指数为 $m=2$,最大循环次数 $I_{\max}=100$,截止误差 $\varepsilon=0.001$,$\alpha=0.3$;

(2)确定初始聚类中心和权值矩阵。从遥感图

像中读入每个像素点的 RGB 值,存入样本集合 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}^{\mathrm{T}}$ 中,取其中 10% 的已知样本信息作为先验知识,进行概率统计,首先得出每一种地物的均值 μ 和标准差 σ,同时通过式(8)计算得出权值矩阵 $\boldsymbol{W}$;然后再利用式(13)和式(14)对先验样本点进行计算,如果需要进行调整,则利用式(16)和式(17)进行聚类中心的调整,每个样本点调整结束之后,即得到了最终的初始聚类中心集合 $\boldsymbol{V}^{(0)}=\{\boldsymbol{v}_1^{(0)},\boldsymbol{v}_2^{(0)},\cdots,\boldsymbol{v}_C^{(0)}\}^{\mathrm{T}}$,开始进行迭代循环计算;

(3)在整个循环过程中,首先加入了排除样本与聚类中心重合的框架,这就避免了隶属度(式(18))会出现无意义的情况;

(4) 通过式(18)和式(19)分别迭代更新隶属度矩阵和聚类中心。式(18)由于加入了权值矩阵 $\boldsymbol{W}$,即对每个数据属性赋予不同的权值,所以更有利于算法的聚类;

(5) 判断两次隶属度矩阵的变化值是否小于截止误差 ε,即 $|\boldsymbol{U}^{(I+1)}-\boldsymbol{U}^{(I)}|<\varepsilon$,或者超过最大循环次数,如果否,则返回步骤(3);否则,算法收敛,执行步骤(6);

(6)最后,经过反模糊化,即可得到每个类别所属于相应地物的样本集合,然后再进行颜色匹配即可得到聚类之后的遥感图像,算法结束。

隶属度计算公式为

$$u_{i,k}^{(I+1)}=\frac{1}{\sum_{j=1}^{c}\left(\frac{[\boldsymbol{x}_k-\boldsymbol{v}_i^{(I)}]^{\mathrm{T}}\boldsymbol{W}[\boldsymbol{x}_k-\boldsymbol{v}_i^{(I)}]}{[\boldsymbol{x}_k-\boldsymbol{v}_j^{(I)}]^{\mathrm{T}}\boldsymbol{W}[\boldsymbol{x}_k-\boldsymbol{v}_j^{(I)}]}\right)^{\frac{1}{m-1}}} \tag{18}$$

聚类中心计算公式为

$$\boldsymbol{v}_i^{(I+1)}=\frac{\sum_{k=1}^{N}(u_{i,k}^{(I)})^m\boldsymbol{x}_k}{\sum_{k=1}^{N}(u_{i,k}^{(I)})^m}\qquad 1\leqslant i\leqslant C \tag{19}$$

4　实验仿真

为验证本文 SWFCM 算法的聚类效果,采用 UCI 数据集和扎龙湿地遥感数据进行了仿真实验,并与其他几种 FCM 算法的聚类效果进行了比较。实验时,将建立起的权值矩阵加入到前期的单点调整算法,先得到邻近于实际类别中心的初始聚类中心,并且将权值加入到隶属度公式的计算中,以更有利于样本点的有效聚类,继而修改循环迭代框架,使 SWFCM 算法可以应用于实际的遥感图像的聚类问题。

4.1　UCI 数据

为了说明本文提出的 SWFCM 算法的真实有效性,首先对机器学习领域著名的 UCI 数据集的 IRIS,Image Segmentation,Wine 和 Pima 等 4 组数据进行了分析和比较,表 1 描述了数据集所包含的类别数目、每个数据所具有的维数和样本集的样本总数,而且每类样本的数目都是相等的;然后,选取 10% 数据作为已知样本信息,其余作为测试样本;最后与标准的 FCM 算法、UFCM[8] 算法的聚类结果进行比较,聚类精度和收敛时间比较结果如表 2 和表 3 所示。

表 1　UCI 数据集

Tab. 1　UCI data sets

数据集	IRIS	Image Segmentation	Wine	Pima
类别数目	3	7	3	2
属性维数	4	19	13	8
数据总数	150	2310	178	768

表 2　UCI 数据集 3 种方法的聚类精度

Tab. 2　Cluster quality of three algorithm on UCI data sets

算法	聚类精度(%)			
	IRIS	Image Segmentation	Wine	Pima
标准 FCM	88.67	61.69	96.07	87.19
UFCM[8]	88.67	54.03	96.07	87.19
SWFCM	91.33	73.33	96.63	92.14

表 3　UCI 数据集 3 种方法的收敛时间

Tab. 3　Convergence time of four algorithm on UCI data sets

算法	收敛时间(ms)			
	IRIS	Image Segmentation	Wine	Pima
标准 FCM	136	169 687	156.25	378.125
UFCM[8]	140.625	107 359	140.625	375
SWFCM	46.875	33 093	125	437.5

从表 2 和表 3 可以发现,用本文提出的 SWFCM 算法对 IRIS,Wine 和 Pima 数据集进行聚类,取得了

较高的聚类精度,其中对于大样本数据集 Image,其聚类精度提高得更加明显,比标准的 FCM 算法提高了 11.64%,比 UFCM 算法提高了 19.3%,而收敛时间却减少了 5~7 倍,同时还可以发现,随着数据维数的增加,精度提高的效果越明显。UFCM 算法采用的是整个样本数据点的均值分布情况,没有考虑具体类别的分布,虽然算法的计算时间减少了,但是聚类的精度却没有提高,反而会有所减小。SWFCM 算法所需要的收敛时间明显减少,但是其中对 Pima 数据进行聚类的收敛时间之所以较大,是因为其中只有一个属性最有利于最终的聚类,应赋予较大的权值的缘故,由于聚类过程完全由这个属性决定,而忽略了其他数据属性的影响,因此需要较多的迭代循环次数才能达到收敛。

表 4 以 IRIS 数据为例,说明了初始聚类中心和最终聚类中心之间的关系。从表中可以看出,由于 UFCM[8] 算法没有考虑实际的样本分布情况,使归一化之后的聚类中心均匀分布在[0,1]之间,所以需要较多的迭代次数才能达到收敛。而本文提出的 SWFCM 算法,由于选取了与最终聚类中心十分邻近的初始点,所以很快达到了收敛,并具有较高聚类精度。

表 4 IRIS 数据集初始聚类中心和最终聚类中心的关系

Tab. 4 The relationship between the initial clustering centers and the final ones for IRIS data sets

算法	初始聚类中心	最终聚类中心	迭代步数	收敛时间(ms)	聚类精度(%)
标准 FCM	{0.499 9,0.416 7,0.661 0,0.708 3} {0.694 4,0.333 3,0.644 0,0.541 7} {0.527 8,0.375 0,0.559 3,0.499 0}	{0.196 0,0.591 1,0.079 1,0.060 0} {0.430 8,0.297 7,0.566 5,0.531 8} {0.702 9,0.452 3,0.795 2,0.827 2}	30	136	88.67
UFCM[8]	{0,0,0,0} {0.333 3,0.333 3,0.333 3,0.333 3} {0.666 7,0.666 7,0.666 7,0.666 7}	{0.196 0,0.591 1,0.079 1,0.060 0} {0.430 8,0.297 7,0.566 5,0.531 8} {0.702 9,0.452 3,0.795 2,0.827 2}	34	140.625	88.67
本文方法	{0.196 1,0.590 8,0.078 6,0.059 9} {0.460 0,0.317 2,0.569 7,0.525 6} {0.657 6,0.424 4,0.784 4,0.830 4}	{0.196 1,0.590 7,0.078 7,0.060 0} {0.453 0,0.310 3,0.559 2,0.510 7} {0.655 2,0.426 1,0.780 5,0.826 5}	18	46.875	91.33

本文以 IRIS 数据为例来说明 3 种方法在迭代过程中的收敛误差,其收敛误差变化曲线如图 2 所示。为了说明加入属性权值和初始聚类中心选择两部分的改进对聚类过程的影响,还与仅用前期单点逼近求取初始聚类中心的模糊 C 均值方法(SFCM)进行了比较。

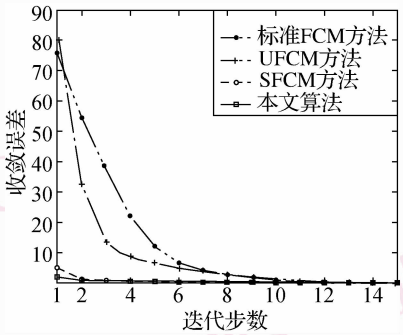

图 2 IRIS 数据集收敛误差变化的比较

Fig. 2 The comparison of convergence error variation for IRIS data set

从图 2 收敛误差的变化曲线可以看出标准 FCM 和 UFCM 方法起始的收敛误差都很大,这是因为初始聚类中心选取得不好,从而导致需要进行多次的循环迭代才可以达到收敛。因为 SFCM 方法是利用单点调整选取了较好的聚类中心,所以初始误差较小,其虽加快了迭代的速度,但是聚类精度与标准的 FCM 方法相同,没有提高。而加入了权值设定的 SWFCM 算法,其速度又得到了进一步的提高,而且最终的聚类精度也提高了。因此,单点调整可更多地加快原始算法的收敛速度,而加入权值则可以提高最终的聚类精度。

4.2 遥感数据

本文以位于黑龙江省西部齐齐哈尔市东南的扎龙湿地自然保护区为研究对象,其地理位置为北纬 46°52′~47°32′,东经 123°47′~124°37′,面积为 2 100 km^2。实验选用 2001 年 10 月 21 日 Landsat TM 多光谱遥感图像,图像分辨率为 30×30 m,截取其中 256×256 大小的图像。首先选取 7 个波段中的 TM2,TM3 和 TM4 通过彩色合成得到了伪彩色遥感图像(见图 3);然后通过对遥感图像进行人工解译,共分耕地、草地、芦苇沼泽、明水沼泽、盐碱地、水体、火烧区和荒地 8 类地物。

图 4 分别是用标准 FCM 算法、UFCM 算法[8]、CFCM 算法[9]和本文提出的 SWFCM 算法对扎龙湿地遥感图像上的各地物进行聚类的结果。观察该聚类图像可以发现,SWFCM 算法可更好地将耕地、草地和芦苇沼泽区分开来,且具有较好的精确程度,对水体附近的盐碱地也有较好的区分。

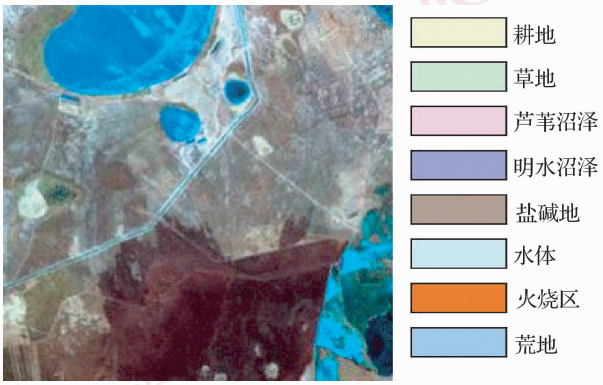

图 3　扎龙湿地原始伪彩色遥感图像

Fig. 3　The original remote sensing image of Zhalong

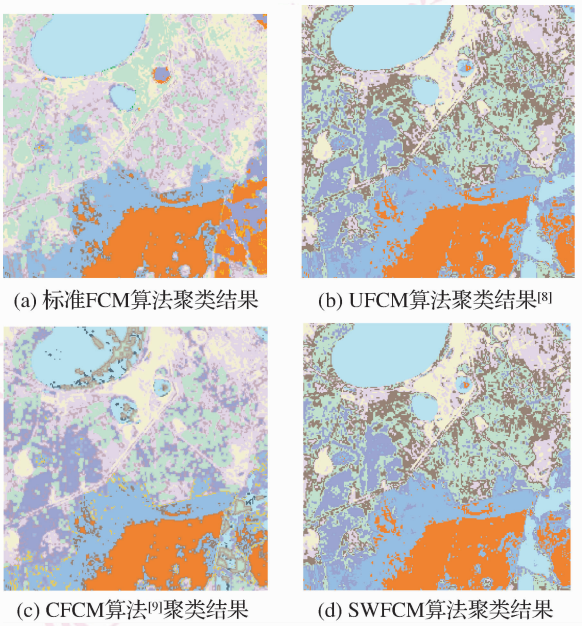

图 4　4 种 FCM 算法的聚类结果图

Fig. 4　The clustering images generated by four FCM algorithms

本文选用误差矩阵的评价方法,其中整体精度代表了整幅遥感图像中各个地物的总体聚类精度,同时辅助以生产者精度来进行多指标综合评价,生产者精度为某类别正确分类的个数除以该类的总采样数。此外,还选用 Kappa 系数作为测量分类数据和参考数据之间一致性的指标,用于比较不同分类器的误差矩阵在精度上的差异,其计算式为

$$Kappa = \frac{n\sum_{k=1}^{q} n_{kk} - \sum_{k=1}^{q} n_{k+}n_{+k}}{n^2 - \sum_{k=1}^{q} n_{k+}n_{+k}} \tag{20}$$

其中,n 为总的样本数目,q 为聚类数,n_{kk} 为第 i 类被正确分为 i 类的数量(一般指误差矩阵的对角线元素),n_{k+} 和 n_{+k} 分别为 i 类的样本数目和被分为第 i 类的数量。

表 5 和表 6 分别列出不同算法聚类的生产者精度、整体精度和 Kappa 系数的比较。由此可以发现,本文提出的 SWFCM 算法,对耕地、草地、芦苇沼泽等光谱值相近的类别具有很好的区分度,其整体精度也比基本 FCM 算法提高了 13.54%,Kappa 系数也从0.686 5提高到了0.841 2,相比 UFCM 和 CFCM 算法,聚类精度也得到了较大的提高。

表 5　生产者精度的比较

Tab. 5　The comparison of producer accuracy

地物类别	不同算法聚类精度(%)			
	FCM	UFCM[8]	CFCM[9]	SWFCM
耕地	80.27	85.98	81.32	98.03
草地	43.53	54.06	45.44	65.54
芦苇沼泽	86.82	70.41	83.55	90.72
明水沼泽	19.73	88.64	89.89	87.02
盐碱地	90.80	90.80	90.80	91.62
水体	81.89	89.44	91.76	96.98
荒地	61.58	79.49	78.38	77.22
火烧区	95.50	79.47	94.37	100.00

表 6　Kappa 系数和整体精度的比较

Tab. 6　The comparison of Kappa and overall accuracy

算法	FCM	UFCM[8]	CFCM[9]	SWFCM
整体精度(%)	77.38	83.29	81.56	90.92
Kappa 系数	0.686 5	0.755 4	0.736 4	0.841 2

5　结　论

本文提出了一种改进的单点逼近型加权模糊 C 均值算法(SWFCM),并以机器学习的典型 UCI 数据集和扎龙湿地遥感数据为例进行了实验分析。SWFCM 算法可以有效地采用部分已知地面真实地

理信息来建立数据属性的权重,并调整原始数据分布,同时利用单点调整算法得到接近于实际类别中心的初始聚类中心,其不仅缩短了运算时间,而且提高了聚类结果的整体精度和 Kappa 系数,可在一定程度上避免陷入局部最小,从而可得到全局最优解。通过与其他算法的比较仿真可以证明,SWFCM 算法克服了数据原始分布特点和初始聚类中心的影响,不仅具有较高的聚类精确度,而且加快了算法的收敛速度,并在进行实际的遥感图像地物聚类的处理过程中得到了很好的应用。

参考文献(References)

1 Tong Qing-xi, Zhang Bing, Zheng Lan-fen. High Spectrum Remote Sensing-Principle, Technology and Application[M]. Beijing: High Education Press, 2006: 184-192. [童庆禧,张兵,郑兰芬. 高光谱遥感—原理、技术与应用[M]. 北京: 高等教育出版社, 2006: 184-192.]

2 Ledoux J. Filtering and the EM-algorithm for the Markovian arrival process[J]. Communications in Statistics-theory and Methods, 2007, **36**(14): 2577-2593.

3 Cannon R L, Dave J V, Bezdek J C, *et al*. Segmentation of a thematic mapper image using the fuzzy C-means clustering algorithm [J]. IEEE Transactions on Geoscience and Remote Sensing, 1986, **GE-24**(3): 400-407.

4 Wen Y C, Isabelle C. Modified fuzzy C-means classification technique for mapping vague wetlands using Landsat ETM + imagery [J]. Hydrological Processes, 2006, **20**(17): 3623-3634.

5 Gath I, Geva A B. Unsupervised optimal fuzzy clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1989, **11**(7): 773-781.

6 Lin Kai-yan, Xu Li-hong, Wu Jun-hui. A fast fuzzy C-means clustering for color image segmentation[J]. Journal of Image and Graphics, 2004, **9**(2): 159-164.[林开颜, 徐立鸿, 吴军辉. 快速模糊 C 均值彩色图像分割方法[J]. 中国图象图形学报, 2004, **9**(2): 159-164.]

7 Yager R R, Filev D P. Approximate clustering via the mountain method [J]. IEEE Transactions on System, Man and Cybernetic, 1994, **24** (8): 1279-1284.

8 Zhang Wen-jun, Gu Xing-fa, Chen Liang-fu, *et al*. An algorithm for initializing of K-means clustering based on mean-standard deviation [J]. Journal of Remote Sensing, 2006, **10**(5): 715-721.[张文君, 顾行发,陈良富等. 基于均值-标准差的 K 均值初始聚类中心选择算法[J]. 遥感学报, 2006,**10**(5):715-719.]

9 Kim Dae-won, Lee KWang H, Lee Doheon. A novel initialization scheme for the fuzzy C-means algorithm for color clustering[J]. Pattern Recognition Letters, 2004, **25**(2):227-237.

10 Wang X Z, Wang Y D, Wang L J. Improving fuzzy C-means clustering based on feature-weight learning[J]. Pattern Recognition Letters, 2004, **25**(10): 1123-1132.

11 Nock R, Nielsen F. On weighting clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, **28** (8): 1223-1235.

12 Pal N R, Bezdek J C. On cluster validity for the fuzzy C-means Model[J]. IEEE Transactions on Fuzzy Systems, 1995, **3**(3): 370-379.

13 Deng Y X, Wilson J P, Sheng J. Effects of variable attribute weights on landform classification [J]. Earth Surface Processes and Landforms, 2006,**31**(11):1452-1462.