环 境 遥 感
第7卷 第3期 REMOTE SENSING OF ENVIRONMENT Vol. 7, No. 3
1992 年 8 月 CHINA Aug., 1992

分级模糊聚类算法及其在遥感积雪识别中的应用

赵帅锋 罗杰波 刘政凯
(中国科学技术大学无线电电子学系信息处理中心)

1991 年 8 月 12 日收稿

摘 要

本文提出了一种分级实现的模糊聚类算法。CFCM 算法具有良好的分类精度，但其初值的选取却是非常困难的。本文所给算法第一级采用改进的 SFCM 算法，其结果作为第二级聚类的初值；第二级采用 CFCM 算法细分。在遥感积雪识别中的实验结果表明，这种算法改善了分类精度，而且由于初值选取较为合理，并不降低分类速度。

关键词 模糊 聚类 积雪 遥感识别

一、引 言

统计模式识别方法可以分为监督与非监督两大类。监督分类方法需要已知所属类别的训练样本集，这常常是办不到的或者代价极其昂贵；非监督分类不需要样本集的先验知识，它按照样本间"距离"亲疏远近的程度，依"物以类聚"的原则，进行分类。所以，研究非监督分类的方法有相当重要的意义。

非监督分类方法中最典型的是动态聚类法，该方法首先选取初始类别中心，然后根据某种聚类准则使各样本向相应的中心聚类，形成新的聚类中心，反复进行上述迭代过程，直到满足误差要求。动态聚类分析中常用的有 K-MEANS 算法和 ISODATA 算法。K-MEANS 算法的聚类效果受预先所选的初始类别数目和初始中心的影响，而 ISODATA 算法的收敛性得不到保证。

由于实际问题的模糊特性，一个样本不是确定地属于某一类，而是不同程度地属于所有类。把模糊集合理论与普通聚类分析相结合，可以得到一个新的聚类分析方法，即模糊聚类。

一般地说，非监督分类的精度不会比监督分类更高，提高非监督分类的分类精度以和监督分法相媲美，是一个值得探索的问题。本文研究探讨了 CFCM 算法初值对结果的影响，形成了分级实现模糊聚类算法。

二、CFCM 聚类算法简介

模糊聚类方法是 E. Ruspini 于 1969 年首先提出的。比较实用的是 J. C. Bezdek 的

FCM (Fuzzy C-Means) 算法[1]，FCM 算法最适合于类别样本呈超球形分布的情况。CFCM 算法在 FCM 算法的基础上，引入了内积诱导矩阵的概念，该矩阵代表各类数据的分布特性，在动态聚类过程中，随着分类情况的变化而变化，最终将尽量符合原始数据的几何拓扑形状，这种算法有较高的分类精度。这两种方法详见文献[2]，下面对 CFCM 原理作一简单介绍。

设 R 为全体实数集，R^p 为 P 维全体实数集，R^+ 为非负全体实数集，W_{cn} 为维数是 $c \times n$ 的全体实数矩阵集。对于多波段遥感图像数据，波段可用一个 P 维向量表示，即

$\bar{X} = (x_1, x_2, \cdots, x_p)$，$\bar{X}$ 的所有集合称为集合 $\boldsymbol{X}$。PD^c 为 c 元正定矩阵向量集。

给定一个有限集 $X \in R^p$，和一个整数 c，$2 \leqslant c \leqslant n$，$X$ 的模糊 c 分割空间定义为集合

$$M_{fc} = \left\{U \in W_{cn} \mid u_{ik} \in [0, 1], \forall i, k; \sum_{i=1}^{c} u_{ik} = 1, \forall k; 0 < \sum_{k=1}^{n} u_{ik} < n, \forall i\right\}$$

CFCM 聚类方法是根据最小方差准则施于类内距离和泛函上得到的。

CFCM 聚类方法的泛函 J_m：$M_{fc} \times R^{cp} \times PD^c \to R^+$ 的定义为：

$$J_m(U, V, A) = \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik}^m \|\vec{x_k} - \vec{V_i}\|_{A_i}^2 \tag{1}$$

其中 $u \in M_{fc}$，M_{fc} 即是 X 的一个模糊 C 分割；$V = (\vec{V_1}, \vec{V_2}, \cdots, \vec{V_c}) \in R^{cp}$，$\vec{V_i} \in R^p$ 是第 i 个聚类中心，$1 \leqslant i \leqslant c$，$m \in (1, \infty)$ 为权指数，u_{ik} 是第 i 个数据对第 k 类的隶属函数。

$\|\vec{x_k} - \vec{V_i}\|_{A_i}^2 = (\vec{x_k} - \vec{V_i})^T A_i (\vec{x_k} - \vec{V_i})$ 为马氏距离，A_i 为正定矩阵，对应于每一类。$A = (A_1, A_2, \cdots, A_c)$ 为矩阵向量。A_i 称为“内积诱导矩阵”，引入 A_i 的思想是：在每一个数据集中，不同的类具有不同的几何形状，因而在聚类准则中所采用的距离范数形式也应有所不同。因此，对每一类均有一相应的相关性对称正定矩阵 A_i。

式(1)中设 $\|\cdot\|$ 为固定范数，固定 $m \in (1, \infty)$，设 X 至少有 $C < N$ 可分类，对任意 K，定义集合：

$$I_k = \{1 \leqslant i \leqslant c, d_{ik} = \|\vec{x_k} - \vec{V_i}\| = 0\}$$

$$\tilde{I_k} = \{1, 2, \cdots, c\} - I_k$$

则 $(u, V) \in M_{fc} \times R^{cp}$，只有当满足下列条件时才能使泛函 J_m 达到整体极小。

(1) 如果 $I_k = \Phi$，则

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}}\right)^{2/(m-1)}} \tag{2}$$

否则，如果 $I_k \neq \Phi$，则 $u_{ik} = 0$，对于任意 $i \in \tilde{I_k}$，而且 $\sum_{i \in I_k} u_{ik} = 1$

(2) 各类别中心矢量为：

$$V_i = \frac{\sum_{k=1}^{n} (u_{ik})^m \vec{x_k}}{\sum_{k=1}^{n} (u_{ik})^m} \tag{3}$$

对任意 i.

设泛函 g: $PD^c \to R^+$, PD^c 为 c 维正定矩阵向量集,则:

$$g(A) = J_m(U, V, A)$$

(u,V)固定,且满足定理 (1), 对于每一个 i, $|A_i| = \sigma_i$ 固定, 则 $g(A)$ 对于 A 的局部极小的必要条件是:

$$A_i = [\sigma_i |T_{fi}|]^{1/p}((T_{fi}^{-1})), \quad 1 \leqslant i \leqslant c \tag{4}$$

其中:

$$T_{fi} = \sum_{k=1}^{n} u_{ik}^m (\vec{x_k} - \vec{V_i})(\vec{x_k} - \vec{V_i})^T \tag{5}$$

当 $A_i = I$ (I 为单位阵), $i = 1,2,\cdots, c$ 时, CFCM 便变为 FCM 聚类方法。

三、SFCM 算 法

SFCM 是一种半模糊聚类算法,它基于这样的模型:把每个样本的模糊性限制在所有聚类中心的一个子集内。这一想法是有实际考虑的,例如,对于一个由两个波段 a 和 b 的图像构成的样本集,每一波段有三个中心值, $a_1 < a_2 < a_3$, $b_1 < b_2 < b_3$, 由此可构成九个可能的聚类中心 (a_1,b_1), $(a_2,b_2),\cdots,(a_3,b_3)$, 但对于任一样本 (a_0,b_0), a_0 不等于 a_1, a_2, a_3, b_0 不等于 b_1, b_2, b_3, 则样本最多可能属于九个类中的四类,即在这四个类上才有非零的隶属度。可以设样本最多属于 t_c 个子集, $1 \leqslant t_c \leqslant c$, 由于 SFCM 算法采取了半模糊的思想,所以初始中心的选取对聚类结果的影响势必较大,一个本属于 i 类的样本 k, 可能因远离初始中心而使初次迭代后便得 $u_{ik} = 0$ 而以后不再加强。 因此可以先设 $t_c = c$, 再根据迭代收敛的情况逐步减小,在保证分类速度的前提下,使精度适当提高。

SFCM 算法如下:

(1) 固定 c 和 m, $2 \leqslant c \leqslant n$, $m \in (1, \infty)$, 初始化 $\{\vec{V}_i^{(0)}\}$, $t_c = c$;

(2) 在第 b 次迭代, $b = 0,1,2,\cdots$;

(3) 依式(2)式和 $\vec{V}^{(b)}$ 更新 $u^{(b)}$ 为 u';

(4) 利用 t_c 更新 u' 为 $u^{(b+1)}$, 对每一个样本 k, $1 \leqslant k \leqslant n$, 给定 stn 和 step。

(a) 选择 $u'_{k(1)}$, $u'_{k(2)},\cdots, u'_{k(t_c)}$ 为集合 $\{u'_{ik}\}$ 中最大的 t_c 个值,取

$$\alpha_k = \min_{1 \leqslant i \leqslant t_c} u'_{k(i)},$$

若 $u'_{ik} < \alpha_k$,

注: stn 为减类误差门限,是逐步减类的一个依据,它随着迭代收敛的情况减小,当然 stn$>\varepsilon$。step 为 stn 减小的步长。

则置 $u'_{ik}=0$，$1\leqslant i\leqslant c$。

(b) 归一化隶属函数：

$$u_{ik}^{(b)}=u'_{ik}/\sum_{j=1}^{c}u'_{jk}$$

(5) 依式(3)和 $u^{(b+1)}$ 计算 c 个模糊聚类中心 $\overrightarrow{V_i^{(b+1)}}$；

(6) 比较 $\overrightarrow{V_i^{(b)}}$ 和 $\overrightarrow{V_i^{(b+1)}}$，如果 $\max\limits_{\substack{1\leqslant i\leqslant c\\1\leqslant j\leqslant p}}\{|V_{ij}^{(b)}-V_{ij}^{(b+1)}|\}>\mathrm{stn}$，则 $b=b+1$，转(2)，否则 (a) 如果 $\max\limits_{\substack{1\leqslant i\leqslant c\\1\leqslant j\leqslant p}}\{|V_{ij}^{(b)}-V_{ij}^{(b+1)}|\}<\varepsilon$，或者 $t_c=1$，停止。

(b) 否则，$b=b+1$，$\mathrm{stn}=\mathrm{stn}-\mathrm{step}$，$t_c=t_c-1$ 转(2)。

四、分级聚类算法

SFCM 算法运算速度较高亦具有相当高的分类精度，特别是分类性能受初始中心的影响不大，所以可以用于聚类算法的初始分类。CFCM 算法具有良好的分类精度，然而其初值的选择无任何先验知识，并且分类速度受初始中心影响，而分类精度则受 $\sigma_i(\|A_i\|=\sigma_i)$ 的影响，所以还无法把 CFCM 算法完全应用于实际工作中。但把二者相结合，便可得到较好的聚类方法。

CFCM 算法的初值包括初始中心和 $\sigma_i(i=1,2,\cdots,c)$ 值。对于每一类而言，σ_i 越大，则样本到该类中心的距离越大，分到该类样本数相应减少[2]，CFCM 完成的粗分类可以求得每一类均方差，均方差值越大，说明此类中含有非本类样本的可能性愈大，这样把粗分类结果所得的每一类均方差值赋 i，使得 CFCM 算法减少此类别数目，而首先减去的样本可能就是那些非本类点，从而使 CFCM 算法完成细分类，并成功地解决了 σ_i 值的选取，同时，粗分类所得聚类中心和由此求得的协方差阵亦可一并作为 CFCM 的初始聚类中心和初始 T_{fi}（求得初始 A_{fi}）。算法的过程如下：

(1) 输入必要参数进行改进 SFCM 聚类；

(2) 依改进 SFCM 结果，求得协方差矩阵 C_{fi}，均方差值 σ_i（每一维方向上的），$i=1,2,\cdots,p$；

(3) 取 $T_{fi}^{(0)}=C_{fi}$，$\sigma_i=[\sigma_{1i}\sigma_{2i}\cdots\sigma_{pi}]^{1/p}$ $(i=1,2,\cdots,c)$，且 $V_{cfcm}^{(0)}=V_{sfcm}$；

(4) 进行 CFCM 聚类；

(5) 输出必要结果。

五、遥感图像处理中的聚类算法调整

在用以上算法做遥感图像处理时，考虑图像像素之间的空间相关性信息，即对某一点而言，其空间邻域像点如果均属于某一类，则此点属于该类的概率就极大。对算法作如下调整：

设 $u_i(j,k)$ 为像点 (j,k) 对于第 i 类的隶属度，又据式(2)按批求得 $u_i(j,k)$ 后，再

调整为:

$$u_i(j,k) = \text{wei} * u_i(j,k) + (1 - \text{wei}) * [u_i(j,k+1) + u_i(j,k-1) + u_i(j-1,k) + u_i(j+1,k)]/4$$

wei 为权重，可适当调整。一般取 wei $\in [0.6, 0.9]$。这样根据周围四邻近像素的隶属度修改该点的隶属度，就在算法中引入了空间相关信息。取 wei $\in [0.6, 0.9]$ 是指像素点自己的隶属度的权重较大，这也是合理的。

六、实验结果

我们采用上述算法，用于遥感图像的分类研究，所用图像为祁连山地区 1984 年 11 月 4 日的 NOAA-AVHRR 遥感图像，图像尺寸为 100 × 100，取一、二、四波段进行分类(共五波段)。实验中取 $c = 4$ (设为四类)，$m = 1.5$，任选初始聚类中心。由于现有的实测手段有限，地面实况无法获得，因此采用目视判读积雪区域，取灰度级上至少有 10 个像点的最大灰度值的 70% 以上的像点为积雪点，共 503 个。实际上目前的分析方法也只是由专家根据经验判读的。表 1 给出分类结果:

表 1 积 雪 分 类

Table 1 Snowpack classification

算 法	初 始 中 心	σ_i	迭代次数	时 间	分类结果	精 度
改进 SFCM	$[0, 0, 0, 0]^T$ $[84, 84, 84, 84]^T$ $[167, 167, 167, 167]^T$ $[250, 250, 250, 250]^T$		12	108 秒	406	81.7%
CFCM	$[0, 0, 0, 0]^T$ $[50, 50, 50, 50]^T$ $[100, 100, 100, 100]^T$ $[150, 150, 150, 150]^T$	$\sigma_1 = 1$ $\sigma_2 = 1$ $\sigma_3 = 1$ $\sigma_4 = 1$	13	198 秒	1011	
分级算法			5+9	146 秒	474	94.3%

以上给出的是积雪点的估计，由于上表中 CFCM 算法所得积雪面积过大，已无统计的意义。分类的效果还可以通过将分类结果图像与遥感波段图像对比，目视观察各种地物的分类效果和层次来看，认为分级算法最好，CFCM 次之，SFCM 最差。

此外，我们还对 TM 图像(日本平琢市区)作了分类实验，TM 图像数据含混度远大于 NOAA 图像，分级聚类算法的正确率仍达到 68.67% (与东京大学测量地面实况比较结果)。

七、结 论

本文提出的分级聚类算法，从 CFCM 算法解决初值问题入手，引入了改进的 SFCM 算法，并由此构成了分级聚类算法。该算法不仅可以保证 CFCM 算法在初值和 σ_i 时的

较快运算速度，而且避免了 CFCM 算法的一些缺点，特别是不好的中心和 σ_i 使得 CFCM 总是处于振荡状态，即收敛速度极慢甚至不收敛。实验结果和理论分析均证明，分级聚类算法是一种很好的模糊聚类方法。我们认为，非监督聚类的性能会比监督聚类的差，但非监督聚类的优点足以弥补这种不足。

参　考　文　献

[1] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, 1981.

[2] 刘政凯、赖海宁、肖锦玉，积雪识别中改进的模糊聚类方法，黄河流域典型地区遥感动态研究，科学出版社，1990。

[3] Shokri, Z. Selim and M. A. Ismail, Soft Clustering of Multidimensional Data: A Semi-Fuzzy Approach, Pattern Recognition, Vol. 17, No. 5, pp. 559—568, 1984.

A HIERARCHICAL FUZZY CLUSTERING ALGORITHM ON SNOW RECOGNITION

Zhao Shuaifeng Luo Jiebo Liu Zhengkai

(Dept. of Radio Electronics, University of Science and Technology of China, Hefei)

Abstract

A hierarchical fuzzy clustering algorithm is presented in this paper. As CFCM algorithm provides good classification resolution but the selecting of initial vectors is of great difficulty and blindness. At the first step, an improved SFCM algorithm is adopted for coarse segmentation and its result serves as initial vector for the fine segmentation by CFCM at the second step. Applying this algorithm to snow recognition shows it can yield very satisfying performance in partition resolution while does not reduce computation speed due to a reasonable initial vector.

Key words　Fuzzy clustering　Snowpack　Remotely sensed recognition