基于模糊 C-均值聚类算法的遥感影像分类

王翠玲，王一冒
中国矿业大学环境与测绘学院，江苏徐州（221008）
E-mail：wangcuiling118@163.com

摘　要：本文简要介绍了 K-均值聚类和模糊 C-均值聚类算法，传统的 k-均值聚类算法广泛用于图像的自动分类,但没有考虑到图像信息的不确定性问题,模糊 C-均值聚类方法有效的解决了遥感信息的不确定性和混合像元的 划分。对徐州影像数据的聚类实验分析表明，模糊 C－均值聚类方法的聚类效果要优于 K－均值聚类方法。

关键词：影像分类，K-均值聚类，模糊 C-均值聚类
中图分类号：TP75

1. 引言

遥感数据的分类分为监督分类和非监督分类，非监督分类也称聚类。聚类就是按照事物间的相似性进行区分和分类的过程，在这一过程中没有教师指导，因此是一种非监督分类。K-均值聚类方法是遥感数据非监督分类最常用的方法之一。K-均值聚类是一种硬划分，它把每个待辨识的像元严格地划分到某个类中，具有非此即彼的性质，而遥感信息的不确定性和混合像元问题使部分像元很难进行非此即彼的划分。实际上遥感数据所反映的大多数地物覆盖在形态和类属方面存在着中介性，没有确定的边界来区分它们。如树林和草地的边界，城乡边界都是渐变的而非确定的。因此考虑各个像元属于各个类别的隶属度，进行软划分才能更好地区分不同地物类别[1]。

模糊 C-均值算法是 Bezdek 提出的著名模糊聚类分析算法。模糊 C-均值聚类是结合模糊集理论和 K-均值聚类提出的，相对于模糊 C-均值算法，K-均值算法也称为硬 C-均值算法。

2. 模糊 C-均值聚类数学基础

模糊 C-均值算法是用模糊集理论对 K-均值算法进行改进，用隶属度来衡量像元属于各个类别的程度。模糊 C-均值算法改变了 K-均值分类中非此即彼的特性。

2.1 硬 C-均值聚类

给定数据集 $X=\{x_1,x_2,...,x_n\}$，其中 $x_k \in R^P$，令 $P(X)$ 为 X 的幂集(集合 X 的所有子集组成的集合)。X 的硬 C-划分就是集合族 $\{A_i \in P(X) | 1 \le i \le c\}$ (c 是正整数)，即 $\{A_1,A_2,...,A_c\}$ 将 X 划分为 C 类，其中每个 A_i 都是一类，并且满足以下两个条件：

$$\bigcup_{i=1}^{c} A_i = X \tag{1}$$

$$A_i \bigcap A_j = \varnothing \quad ,1 \le i \ne j \le c \tag{2}$$

硬 C-划分可以用 A_i 中元素 x_k 的特征(隶属度)函数来描述。x_k 属于 A_i，则 u_{ik} =1； x_k 不属于 A_i，则 u_{ik} =0。因此给定 u_{ik} 的值就可以惟一确定 X 的一个硬划分。在硬 C-分类中像元隶属度 u_{ik} 满足以下三个条件：

$$u_{ik} \in \{0,1\} \quad ,1 \le i \le c, 1 \le k \le n \tag{3}$$

$$\sum_{i=1}^{c} u_{ik} = 1 \quad , k \in \{1, 2, \ldots, n\} \tag{4}$$

$$0 < \sum_{k=1}^{n} u_{ik} < n \quad , i \in \{1, 2, \ldots, c\} \tag{5}$$

用各个像元的隶属度 u_{ik} 构成矩阵 $U_{c\times n}$，即可得到硬 C-划分的矩阵形式。

在硬 C-划分中，令 $X = \{x_1, x_2, \ldots, x_n\}$ 为任意集合，V_{cn} 为实 $c\times n$ 阶矩阵 $U = [u_{ik}]_{c\times n}$ 的集合，U 满足式(3) (4) (5)，c 为大于等于 2 小于 n 的整数，则 X 的硬 C-划分空间为集合 M_c：

$$M_c = \{U \mid U \in V_{cn}\} \tag{6}$$

硬 C-划分存在两个问题：一是像元划分时非此即彼的性质，即隶属度 u_{ik} 只能取 0 或 1，而实际上边界和混合像元不具有严格的非此即彼的性质；二是像元的划分空间过大，造成寻找最佳划分的难度大。

2.2 模糊 C-均值聚类

模糊 C-均值聚类中数据集 X 的划分空间为集合 M_{fc}：

$$M_{fc} = \{U \in V_{cn} \mid u_{ik} \in [0,1], 1 \le i \le c, 1 \le k \le n\} \tag{7}$$

其中 u_{ik} 为像元隶属于类 A_i 的隶属度，并满足：

$$\sum_{i=1}^{c} u_{ik} = 1,\ k \in \{1, 2, \ldots, n\} \tag{8}$$

在硬 C 分类中由于可能划分空间集合很大，寻找最优划分困难。在模糊 C-均值聚类中，将隶属度 u_{ik} 变为一个连续变量使其在[0，1]区间上取值，这样就能算出某个目标函数关于 u_{ik} 的梯度。进而利用梯度找到最好的搜索方向，简化寻找最优划分的方法。

从像元集的可能划分集合中寻找最优划分的方法之一为目标函数法。最常用的目标函数是总体组内误差平方和：

$$J_w(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik} \| x_k - v_i \|^2 \tag{9}$$

式中 $U \in M_{fc}$ 或 M_c， $V = (v_1, v_2, \ldots, v_c)$， v_i 是类 A_i 的中心：

$$v_i = \sum_{k=1}^{n} u_{ik} x_k \Big/ \sum_{k=1}^{n} u_{ik} \tag{10}$$

在目标函数法中，最优的分类就是使目标函数达到局部最小值的分类。如果一类中的所有像元都贴近于它们的聚类中心，则目标函数很小。K-均值聚类算法就是常见的对应硬 C-划分空间的寻找目标函数最小值从而求得最优划分的算法之一。

模糊 C-均值算法实质上是一种局部寻优方法[2]，算法的目标在于找到 $U = [u_{ik}] \in M_{fc}$ 和 $V = (v_1, v_2, \ldots, v_c)$ ($v_i \in R^P$)，使目标函数 $J_m(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{c} (u_{ik})^m \| x_k - v_i \|^2$ 取局部最小值。式中 $m \in (1, \infty)$ 为模糊加权指数。对 u_{ik} 添加模糊加权指数 m，对于从硬分类推广到软分类时给隶属度一个权重。参数 m 又称为平滑因子，控制着模式在模糊类间的分享程度，最佳 m 的选取尚缺乏理论指导。模糊加权指数 m =1 时即为硬分类， m 越大隶属度越平滑，

m 的经验取值范围为 $m \in [1.5, 2.5]$[3]。目标函数取局部最小值的必要条件为：

$$u_{ik} \frac{1}{\sum_{j=1}^{c}(\frac{\| x_k - v_i \|}{\| x_k - v_j \|})^{\frac{2}{m-1}}}, \ 1 \le i \le c, \ 1 \le k \le n \tag{11}$$

$$v_i = \frac{\sum_{k=1}^{n}(u_{ik})^m x_k}{\sum_{k=1}^{n}(u_{ik})^m}, \ 1 \le i \le c \tag{12}$$

模糊 C-均值算法步骤为：

(1)确定聚类数 c，加权系数 m，中止误差 ε，最大迭代次数 $LOOP$。

(2)初始化隶属度矩阵 $U^{(0)}$。

(3)开始循环，当迭代次数为 IT $(IT = 0,1,2,...)$ 时，根据 $U^{(IT)}$ 计算 C-均值向量：

$$v_i^{(IT)} = \sum_{k=1}^{N}(u_{ik})^m x_k \Big/ (\sum_{k=1}^{N}(u_{ik})^m), \quad i = 1,2,...,c \tag{13}$$

(4)对 $k = 1,2,...,N$，更新 $U^{(IT)}$ 为 $U^{(IT+1)}$：

若 $x_k \ne v_i$ 对所有的 $v_i (i = 1,2,...,c)$ 满足，则对此 x_k 计算 u_{ik}：

$$u_{ik} = [\sum_{j=1}^{c}(\frac{d_{ik}}{d_{jk}})^{\frac{2}{m-1}}]^{-1}, \ x_k \ne v_i, \ i = 1,2,...,c \tag{14}$$

若对某一个 v_i，有 x_k 满足 $x_k = v_i$，则对应此 x_k，令 $u_{ik} = 1$；$u_{ik} = 0 (j \ne i)$。这样，把聚类中心与样本一致的情形去除掉，把隶属度模糊化为 0 到 1 之间的实数。

(5)若 $\| U^{(IT)} - U^{(IT+1)} \| < \varepsilon$ 或 $IT > LOOP$ 停止，否则置 $IT = IT + 1$，返回(3)。

模糊 C-均值算法允许自由选取聚类个数，每一向量按其隶属度 $u_{ik} \in [0,1]$ 聚类到每一聚类中心。模糊 C-均值算法是通过最小化目标函数来实现数据聚类的。

3. 实验结果与分析

实验所用数据为 2002 年徐州市影像数据，实验所选研究区域为 512×512 个像元。设置聚类个数为 7，中止误差 ε =0.001，最大迭代次数 $LOOP$ 为 100，加权指数 m 为 2。比较图 2 模糊 C-均值聚类结果和图 3 K-均值聚类结果，原始影像方框内的道路经模糊 C-均值聚类后仍较完整，大的区域内镶嵌的其他类别的细小区域也保留下来了，而 K-均值分类后道路断开连续性较差而且造成小区域丢失。表明模糊 C-均值聚类结果优于 K-均值聚类结果。

图 1 原始影像

Fig1. originality image

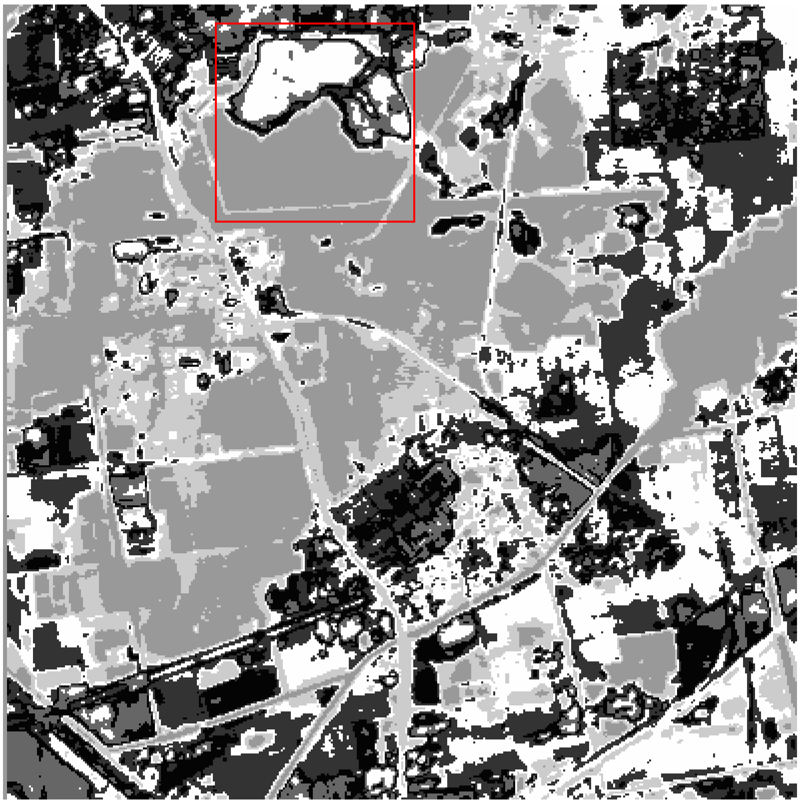

图 2 模糊 C-均值聚类结果

Fig2. fuzzy c-means clustering result

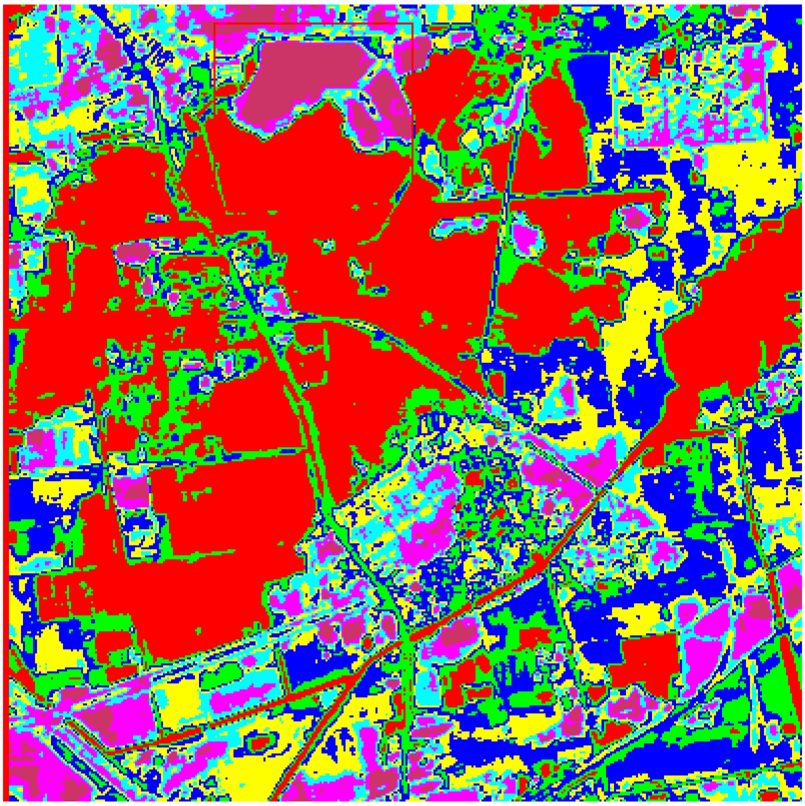

图 3　K-均值聚类结果

Fig3. k-means clustering result

4. 结论

模糊 C-均值聚类考虑遥感影像中信息的不确定性和混合像元问题，计算输入影像中各个像元属于各个类别的隶属度，根据隶属度进行软划分从而更好地区分不同的地物类别。基于模糊集理论而提出的模糊 c－均值聚类可以很好地进行软划分。通过对实验区的遥感数据进行 K－均值聚类、模糊 c－均值聚类。均值聚类的分析表明，模糊 c－均值聚类使得分类后的图像很好地区分了地物类别，取得了优于 K－均值聚类方法的分类结果。

参考文献

[1] 马建文,李启青,哈斯巴干等.遥感数据智能处理方法与程序设计[M]，北京：科学出版社，,2005，112-118
[2] 薛忠,谢维信.模糊 C 均值聚类算法的一种初始化方法[J]，系统工程与电子技术. 1995，11:64-69
[3] 高新波,裴继红,谢维信.模糊 C-均值聚类算法中加权指数 m 的研究[J]，电子学报. 2000，28(4):80-83

Classification Methods of Remote Sensing Image Based on FCM clustering

Wang CuiLing，Wang YiMao
School of Environment Science and Spatial Informatics，China University of Mining and Technology，Xuzhou，Jiangsu　(221008)

Abstract
Conventional k-means clustering algorithm has been widely used in automated image classification, however, it was not successful to classify the Uncertainty information of images. Fuzzy c-means clustering ravels out effectively the compartmentalization of Uncertainty information and Mixed Pixel. Satellite data of Xuzhou is used for testing the results proved that the fuzzy c-mean classifier is more precedence than k-mean classifier.

Keywords：Image Classification；Fuzzy k-means clustering；fuzzy c-means clustering