

基于 K 均值的遥感图像自动识别分类

胡高翔¹, 韩孜²

¹武汉大学遥感信息工程学院, (430079)

²宁波市浙江万里学院计算机与信息学院, (315100)

hgxida@126.com, hz110hz@163.com

摘要: 遥感图像的计算机分类是模式识别技术在遥感技术领域中的具体应用。本文采用了模式识别分类中非监督分类中 k 均值聚类方法对多维遥感图像进行分类,从而达到提取所需地物信息的目的。

关键词: 模式识别、K 均值聚类、遥感影像

1. 引言

遥感图像的计算机分类,就是对地球表面及其环境在遥感图像上的信息进行属性的识别与分类,从而达到识别图像信息所相应的实际地物,提取所需地物信息的目的。遥感图像的计算机分类是模式识别中的一个方面,它的主要是别对象是遥感图像及各种变换后的特征图像。目前,遥感图像的自动识别方法主要采用统计方法,按照决策理论方法,需要从被识别的模式(即对象)中,提取一组反映模式属性的量测值,称之为特征,并把模式特征定义在一个特征空间中,进而利用决策的原理对特征空间进行划分。以区分具有不同特征的模式,达到分类的目的。

统计模式识别从根本上讲,是利用各类的分布特征,即直接利用各类的概率密度函数、后验概率等,或隐含地利用上述概念进行分类识别。其中的聚类分析法是利用待分类模式之间的“相似性”进行分类,较相似的作为一类,较不相似的作为另一类。在分类过程中不断地计算所分划的各类的中心,一个待分类模式与各类中心的距离作为对其分类的依据。这实际上是在某些设定下隐含地利用了概率分布概念,因常见的概率密度函数中,距期望值较近的点概率密度较大。该类方法的另一种技术是根据待分类模式和已指判出类别的模式的距离来确定其类别,这实际上也是在一定程度上利用了有关概念。K 均值聚类方法是统计模式识别分类中非监督分类法的一种。

2. 相关背景知识

2.1 非监督分类

非监督分类(也称聚类分析)算法中并没有显式的信号来对训练样本集中的每个输入样本提供类别标记和分类代价,并寻找能降低总体代价的方向。系统对输入样本自动形成“聚类”或“自然”的组织。所谓“自然”与否是由聚类系统所采用的显式或隐式的准则确定的。给定一个特定的模式集和代价函数,不同的聚类算法将导致不同的结果。通常要求用户事先指定预定的聚类的数目。

遥感影像的非监督分类是指人们事先对分类过程不施加任何的先验知识,而仅凭遥感影

像地物的光谱特征的分布规律，即自然聚类的特性进行“盲目”的分类。

非监督分类的基本思想就是先选择若干个模式点作为聚类中心，每一个中心代表一个类别，按照某种相似性度量方法将各模式归于各聚类中心所代表的类别，形成初始分类。然后由聚类准则判断初始分类是否合理，如果不合理就修改分类，如此反复迭代运算，直至合理为止。

2.2 类的定义与类间距离

2.1.1 类的定义

类的划分具有人为规定性。一个分类的结果优劣最后只能根据实际来评价，因此较多地利用研究对象的知识才能选择适当的类的定义，从而使分类结果更符合实际。

下面给出几组常用的类的定义：

- (1) 集合 S 中任两个元素 x_i 、 x_j 的距离 d_{ij} 有

$$d_{ij} \leq h \quad (2.1.1.1)$$

其中 h 为给定的某个域值，称 S 对于域值 h 组成一类。

- (2) 集合 S 中任两个元素 x_i 、 x_j 的距离 d_{ij} 有

$$\frac{1}{(k-1)} \sum_{x_j \in S} d_{ij} \leq h \quad (2.1.1.2)$$

其中 k 为集合 S 中元素的个数， h 为给定的某个域值，称 S 对于域值 h 组成一类。

- (3) 集合 S 中任两个元素 x_i 、 x_j 的距离 d_{ij} 有

$$\frac{1}{k(k-1)} \sum_{x_i \in S} \sum_{x_j \in S} d_{ij} \leq h$$

$$d_{ij} \leq r \quad (2.1.1.3)$$

其中 k 为集合 S 中元素的个数， h 、 r 为给定的某个域值，称 S 对于域值 h 、 r 组成一类。

2.1.2 类间距离

在有些聚类算法中要用到类间距离，下面给出几个常用类间距离定义式。

- (1) 最近距离

两个聚类 ω_i 和 ω_j 之间最近距离定义为

$$D_{kli} = \min_{i,j} [d_{ij}] \quad (2.1.2.1)$$

式中， d_{ij} 表示 $x_i \in \omega_k$ 和 $x_j \in \omega_l$ 之间的距离。

(2) 最远距离

两个聚类 ω_i 和 ω_j 之间最远距离定义为

$$D_{kli} = \max_{i,j} [d_{ij}] \quad (2.1.2.2)$$

式中, d_{ij} 表示 $x_i \in \omega_k$ 和 $x_j \in \omega_l$ 之间的距离。

(3) 平均距离

两个聚类 ω_p 和 ω_q 之间最近距离定义为

$$D_{pq}^2 = \frac{1}{n_p n_q} \sum_{\substack{x_i \in \omega_p \\ x_j \in \omega_q}} d_{ij}^2 \quad (2.1.2.3)$$

式中, D_{ij} 表示 $x_i \in \omega_k$ 和 $x_j \in \omega_l$ 之间的平均平方距离。

3. k-均值分类算法

3.1 基本思想

K-均值聚类是非监督分类方法中的一种。K-均值算法的聚类准则是使每一个聚类中, 多模式点到该类别的中心距离的平方和最小。其基本思想是, 通过迭代, 逐次移动各类的中心, 直至得到最好的聚类结果为止。

3.2 算法步骤

(1) 假设分类类别数为 m 个, 任选 m 个类的初始中心 $Z_1^{(0)}$ 、 $Z_2^{(0)}$ $Z_m^{(0)}$ 。

(2) 将待分类的模式矢量特征集 $\{X_i\}$ 中的模式逐个按最小距离原则分划给 m 类中的某一类, 即

如果 $d_{il}^{(k)} = \min[d_{ij}^{(k)}], i=1,2,\dots,N$

则判 $x_i \in \omega_l^{(k+1)}$ 。

式中 $d_{ij}^{(k)}$ 表示 x_i 和 $\omega_j^{(k)}$ 的中心 $Z_j^{(k)}$ 的距离, 上角标表示迭代次数。于是产生新的聚类 $\omega_j^{(k+1)}$ ($j=1,2,\dots,m$)。

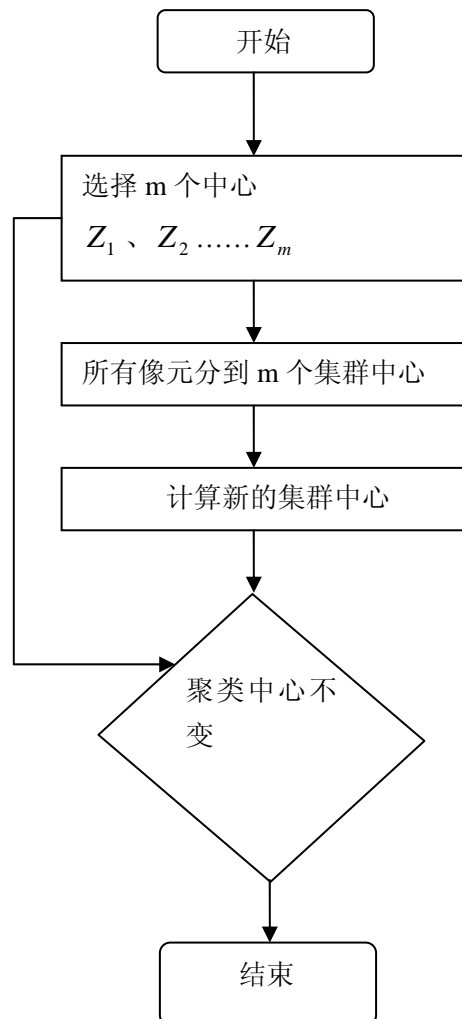
(3) 计算重新分类后的各类心

$$z_j^{(k+1)} = \frac{1}{n_j^{(k+1)}} \sum_{x_i \in \omega_j^{(k+1)}} x_i, \quad j=1,2,\dots,m$$

式中 $n_j^{(k+1)}$ 为 $\omega_j^{(k+1)}$ 类中所含模式的个数。

(4) 如果 $z_j^{(k+1)} = z_j^{(k)}$ ($j=1,2,\dots,m$), 则结束; 否则, $k=k+1$, 转至(2)

3.3 K-均值算法流程图



3.4 K-均值算法表达

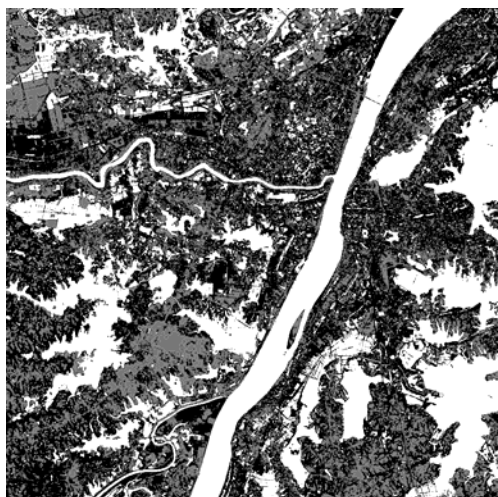
```
1  begin initialize n,m   $Z_1^{(0)}$ 、 $Z_2^{(0)}$  .....  $Z_m^{(0)}$ 
2      do 按照最近邻  $Z_i$  分类 m 个样本
3          重计算  $Z_i$ 
4      until  $Z_i$  不在改变
5  return  $Z_1$ 、 $Z_2$  .....  $Z_m$ 
6  end
```

4. 实验结果

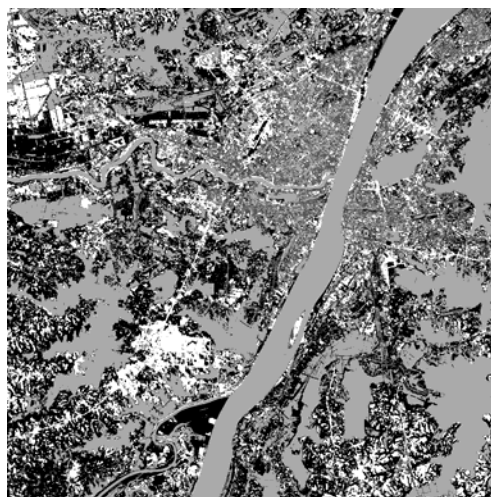
本实验在微机上进行。其微机性能为：Pentium4 CPU 1.6GHz 256MB 内存。在 VC++

环境下，用 MFC 开发了 K-均值遥感图像分类程序，进行有关实验研究。本次研究的数据主要是 TM1,2,3,4,5,7 六个波段的数据。

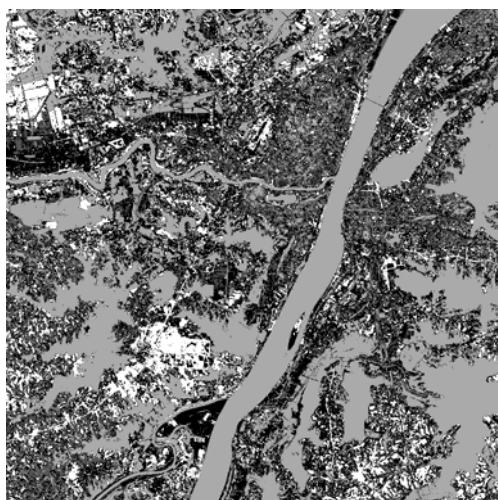
下面是对 TM1,2,3,4,5,7 六个波段遥感图像分类的结果。



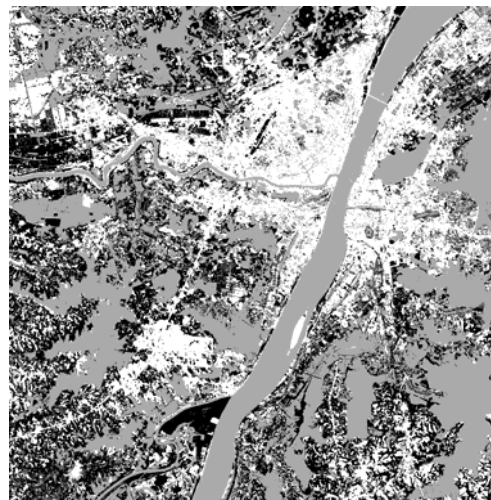
(图 4.1 分类类别数 3，迭代次数 6)



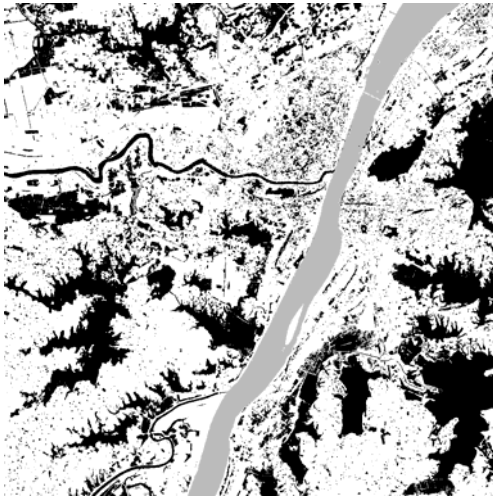
(图 4.2 分类类别数 4，迭代次数 6)



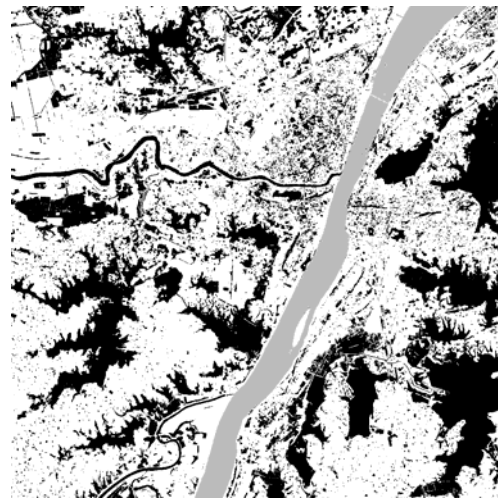
(图 4.3 分类类别数 4，迭代次数 8)



(图 4.4 分类类别数 4，迭代次数 10)



(图 4.5 分类类别数 5, 迭代次数 3)



(图 4.6 分类类别数 5, 迭代次数 6)

由图中我们可以很清楚可以看到, K 均值分类整体效果较好, 然而分类数和迭代次数不同, 分类效果有一定差异。图 4.1 中, 分类类别数为 3, 如图可知, 成功将湖泊, 河流与其他地物分离开来, 因而效果比较好。而图 4.2 中, 分类结果得出湖泊, 河流, 房屋, 土地, 但房屋与土地分离得不是很完美, 相比之下图 4.3 效果比较好, 而图 4.4 迭代次数增多, 并没有使分类效果提高反而房屋与土地的分离效果比较差。同样, 这一点在图 4.5 与图 4.6 的对比上也有所体现。这是和算法本身的原因有关。因为算法结果受到所选聚类中心的数目和其初始位置及模式分布的几何性质和读入次序等因素的影响, 并在迭代过程中没有调整类数的措施, 因此可能产生不同的初始分类得到不同的结果。

5. 结论

K-均值是以确定的类数及选定的初始聚类中心为前提, 使各模式到其所判属类别中心距离之和最小的最佳聚类。方法简单, 结果比较令人满意。但是由于算法受到所选聚类中心的数目和其初始位置以及模式分布的几何性质和读入次序等因素的影响, 并在迭代过程中又没有调整类数的措施, 因此可能产生不同的初始分类得到不同的结果。

为了提高分类结果, 我们可以对初始聚类进行一些处理, 通过一些简单的聚类中心试探方法来找出初始聚类中心而不是单单随机选取初始聚类中心。

参考文献

- [1] 张莉, 孙钢, 郭军 基于 K-均值聚类的无监督的特征选择方法 计算机应用研究 2005 年第 3 期
- [2] 孙家柄等《遥感原理与应用》, 武汉: 武汉大学出版社, 2003.2。
- [3] Richard O.Duda Peter E.Hart David G.Stork 《模式分类》, 北京: 机械工业出版社, 2003.9
- [4] 孙即祥 《现代模式识别》长沙: 国防科技大学出版社, 2002.1

Automatic recognition and classification of RS imagery based on k-means clustering method

Hu Gao-xiang¹, Han Zi²

¹School of Remote Sensing Information Engineering, Wuhan university, 430079

²The Computer Science and Information Technology department, Zhejiang Wanli university, 315100

Abstract

Computer classification based on RS imagery is a general task of pattern recognition in the field of remote sensing. In this paper, in order to extract required features on earth, a method based on k-means, a classical unsupervised classification method in pattern recognition, is proposed to classifier the multi-dimension RS imagery.

Keywords: *pattern recognition, k-means clustering, RS imagery*

作者简介: 胡高翔, 武汉大学遥感信息学院 2005 级研究生
韩孜, 浙江万里学院, 计算机与信息学院教师