Journal of Image and Signal Processing 图像与信号处理, 2018, 7(1), 57-64
Published Online January 2018 in Hans. http://www.hanspub.org/journal/jisp
https://doi.org/10.12677/jisp.2018.71007

Hans汉斯

# Time and Season Recognition Method via K-Means

**Jianwen Chen, Guanlei Xu**

Dalian Navy Academy, Dalian Liaoning
Email: abcd522479982@qq.com

Received: Jan. 7th, 2018; accepted: Jan. 22nd, 2018; published: Jan. 30th, 2018

## Abstract

**Different time seasons have different characteristics. According to the characteristics of the season, we use various methods on the computer to clearly identify different seasons. Based on HSV color model, the characteristics of the images of each season are analyzed and studied, and in the colors contrast, the mean value and variance of them are compared. Through using the optical image processing technology, adopting the nearest neighbor classification method, comparing the characteristics of different season images, and then classifying and recognizing the selected images, thus the season can be identified automatically. The experimental results confirm the feasibility of this method, and can achieve high performance in season recognition.**

## Keywords

**Feature, HSV Image Color Model, Neighborhood Classification, Image Identification**

# 基于K-Means的时间季节反演识别方法

**陈建文，徐冠雷**

大连舰艇学院，辽宁 大连
Email: abcd522479982@qq.com

收稿日期：2018年1月7日；录用日期：2018年1月22日；发布日期：2018年1月30日

## 摘 要

**不同的时间季节有不同的特征，根据季节的特征，我们在计算机上利用各种方法可以清楚的识别不同的季节。本文利用HSV图像色彩模型，主要对各个季节的图片的特征进行分析和研究，进行颜色间的对比，比较其色彩值的平均值与方差，运用光学图像处理技术，采用近邻分类的方法，对比各季节图像的特征**

文章引用: 陈建文, 徐冠雷. 基于 K-Means 的时间季节反演识别方法[J]. 图像与信号处理, 2018, 7(1): 57-64.
DOI: 10.12677/jisp.2018.71007

**差异，对所选图像进行分类与识别，进而达到自动识别季节的目的。实验结果证实了此研究方法的可行性，在季节识别上能够达到自动识别的效果。**

## 关键词

**特征，HSV图像色彩模型，近邻分类，图像识别**



## 1. 引言

随着互联网技术与数字化的飞速发展，以及电子数码产品的普及，人们能够获取的数字图像数据已经越来越多。图像作为一种表现直观、内容丰富的多媒体信息，在各行各业中也得到越来越广泛的应用，如数字多媒体图书馆、医学图像应用管理、卫星遥感图像和地理信息系统、身份认证识别系统、电子商务、商标版权的监管等。随之带来的是图像处理技术的不断提升，目前，图形分类主要包括图像采集、图像预处理、图像特征值的选择和提取、图像分类模型建立四个步骤[1]。图像处理方法的重要性源于两个主要应用领域：1) 改善图示信息以便人们解释；2) 为存储、传输和表示而对图像数据进行处理，以便于机器自动理解。本文基于光学图像处理技术，针对天气现象光学特征来反演时间季节，对不同时间季节图片的天气特征进行了提取并研究，给出合理分类，为天气现象的自动分类和观测提供光学技术、方法和手段。

图像特征提取是自动识别图像的基础。获得高效的图像不仅可以降低计算复杂度，还可以准确地表述图像的原始信息[2]。目前，常用的图像底层特征包括颜色、纹理和形状等。在基于内容的图像处理技术中特征提取有着重要的意义：一方面，图像的特征提取与表示是图像识别过程中一个很重要的步骤，所提取的特征直接影响后续的图像相似度匹配过程，识别精度等；另一方面，由于图像所包含的内容和图像类型各异，有些图像颜色特征较为突出，有些图像纹理特征占主导，有些图像则可能用形状来描述更加合适，而有些图像单纯用某一种特征都不足以表示，因此我们需要研究新颖全面的方法来提取图像特征并表示图像。此外，新颖合适的图像特征提取方法对别的领域如人工智能、模式识别等也会有一定的促进作用[3]。因此，如何有效的提取低层视觉特征成为基于内容的图像识别中的一个关键步骤，也成为图像识别领域研究的热点内容。

在前人的研究基础上加上教员的指导下，本文基于 HSV 的颜色空间模型，提取了不同季节图像中的 H (色调)、S (饱和度)、V (亮度)的数据，并对这三类数据进行了求平均、方差，最后再利用 K-means 聚类算法等方法对之前提到的数据进行了分类聚类研究，最终达到时间季节自动识别效果。本文着重于使用 matlab 对上述提出的算法进行时间分析并设计了图像识别的有效程序。

## 2. 图像分类

### 2.1. 典型的分类方法

1) 神经网络分类[4]的方法

人工神经网络是人们对自身大脑神经网络认识理解的基础上人工构造的可以实现某种功能的网络系

统[5]。人工神经网络是对生物神经元的模拟，生物的每一个神经元相当于一个基本单元，且每个单元关联一个权值。神经网络较复杂，需要较长的时间来训练，因而这种方法比较适合训练时间长的应用。神经网络的参数大多数靠总结历史数据所得的经验所得。一般可分为以下几类：

人工神经网络从结构方面讲可以分为前馈网络(如 BP 网络)和反馈网络(如 Hopfield)；从状态方面来讲又可分为离散性网络和连续性网络；从学习的方法方面又可分为监督学习网络(如 BP、RBF 网络)和无监督学习网络[6] (如自组织网络)。

2) 贝叶斯分类方法

这个分类方法的主要思想：先计算每个类别在训练集的分布，将这种分布看作为不同类别的概率分布，最后在测试过程中运用统计学原理和概率统计原理中的贝叶斯定理来估计某个特定样本属于某一类的概率 0.6。

目前贝叶斯方法主要包括贝叶斯信念网络和朴素贝叶斯方法两类，其中前者在当前社会应用最多，但是这种方法的训练较为复杂，用来评估的函数比较难选，这是现在需要研究解决的问题。

## 2.2. 图像聚类的研究现状及发展

聚类分析作为数据挖掘领域的一个重要分支，已被广泛研究多年。当人们利用数据挖掘工具对数据关系和模型进行识别的时候，通常首要步骤就是聚类，目的是将初始时大量无规律的数据按照一定的规则重新组合成若干类，使同一类内各对象之间尽可能最相似、不同类之间尽可能不相似，以揭示数据分布的规律性，发现数据属性之间重要的相互关系以及数据全局的分布模式[7]。聚类分析在客户分类、基因识别、文本分类、图像处理中有着广泛的应用[8]。

聚类分析是个具有很强挑战性的领域，目前的研究主要有以下几个方向：

1) 发现任意形状的聚类的能力

许多聚类分析算法采用的是基于欧氏距离或曼哈顿距离的相似度度量方法来决定簇，这类算法通常趋于发现的是一些尺寸和密度相近的、球状的类[9]。但是，现实数据库中的聚类可以是任意的形状，甚至是具有分层树的形状，故要求算法具有发现任意形状的聚类的能力。

2) 输入参数对领域知识的弱依赖性

一些聚类算法在聚类分析开始前要求用户输入一定的参数，例如期望得到的聚类数目、置信度、支持度等。最终聚类的结果通常对这些参数较为敏感[10]。另一方面，对于高维对象数据集来说，这些参数的值通常也很难确定。这样不仅加重了用户使用该聚类分析工具的负担，也难以保障聚类质量的好坏。因此就需要一个较好的方案来解决这个问题。

3) 初始值的选择以及输入顺序对聚类结果的影响

有些聚类算法对初始值的选择和数据的输入顺序很敏感。即对同一个数据集，将其以不同的次序输入到分析算法中，可能得到差别很大的聚类。开发对输入不敏感的算法是目前研究的一个重点。

4) 高维数据的处理能力

数据库或数据仓库都包含若干字段或者是属性，一些聚类分析算法对处理维数较少的数据集时效果不错，例如二维三维数据[11]。但是对于高维数据空间，比如图像数据、视频数据等就没那么乐观了。所以对高维数据的聚类分析是非常有挑战性的，特别是考虑到数据在高维空间中，其数据分布通常比较稀疏，而且形状也可能不规则，因此这方面的研究是当前聚类分析中一个很重要的分支。

目前，主要存在着如下几种聚类方法:划分方法、层次方法、基于密度的方法、基于网格的方法以及基于模型的方法等。图像聚类的目的是在图像识别前为图像数据库建立有效的识别类型，以加快图像识别速度，提高图像识别的精确性。图像聚类的关键问题之一是如何选取合适高效的聚类算法。目前应用

于图像数据的聚类分析算法主要有 K-means 算法、模糊 C 均值方法、遗传算法、近邻传播算法(Affinity Propagation, AP)以及这些方法的改进等[12]。

## 3. HSV 空间模型

HSV 模型在 1978 年由埃尔维・雷・史密斯创立(图 1)。

HSV 色彩模型从 CIE 三维颜色空间演变而来，它采用的是用户直观的色彩描述方法，它跟孟塞尔显色系统的 HVC 球型色立体较接近。(如图 HSV 色彩六棱锥)只不过 HSV 色彩模型是一个倒立的六菱锥，只相当于孟塞尔球型色立体的一半(南半球)，所以不含黑色的纯净颜色都处于六菱锥顶面的一个色平面上[13]。在 HSV 六菱锥色彩模型中，色相(H)处于平行于六菱锥顶面的色平面上，它们围绕中心轴 V 旋转和变化，红、黄、绿、青、蓝、品红六个标准色分别相隔 60 度[14]。色彩明度(B)沿六菱锥中心轴 V 从上至下变化，中心轴顶端呈白色(V = 1)，底端呈黑色(V = 0)，它们表示无彩色系的灰度颜色。色彩饱和度(S)沿水平方向变化，越接近六菱锥中心轴的色彩，其饱和度越低，六边形正中心的色彩饱和度为零(S = 0)，与最高明度的 V = 1 相重合，最高饱和度的颜色则处于六边形外框的边缘线上(S = 1) [15]。

1) 色相、饱和度与六棱锥色平面(H 和 S)色平面(H、S)的基础是 CIE 色度图的 x、y 色平面。

2) 明度与六棱锥中轴色(v)色明度(V)的基础是 CIE 三维颜色空间的亮度因素 Y。

## 4. 最近邻分类

假定有 $c$ 个类别 $\omega_1,\omega_2,\cdots,\omega_c$，$\omega_c$ 的模式识别问题，每类有标明类别的样本有 $N_i$ 个，那么可以轨道 $\omega_i$ 类的判别函数为 $g_i(x)=\min\left\|x=x_i^k\right\|$，其中 $k=1,2,\cdots,N_i$。$i$ 表示 $\omega_i$ 类，$k$ 表示 $\omega_i$ 类的 $N_i$ 个样本中的第 $k$ 个。决策规则为：若 $g_j(x)=m_k ing_i(x), i=1,2,\cdots,c$，则决策 $x\in\omega_j$ [16]。

最近邻分类的另一个直观解释是:令 $D^n=\{x_1,x_2,\cdots,x_n\}$，其中每一个样本 $x_i$ 所属的类别均已标记。对于参数样本点 $x$，在集合 $D^n$ 中距离它最近的点记为 $x'$ 那么最近邻分类规则就是把点 $x$ 分为 $x'$ 所属的类别[17]。

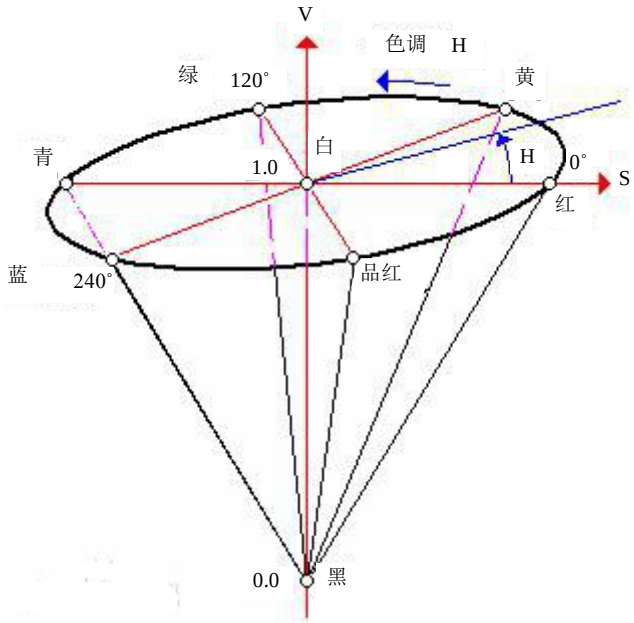


**Figure 1.** HSV model
**图 1.** HSV 模型

## 5. K-Means 聚类算法

K-means 算法是硬聚类算法，是典型的基于原型的目标函数聚类方法的代表，它是数据点到原型的某种距离作为优化的目标函数，利用函数求极值的方法得到迭代运算的调整规则。K-means 算法以欧式距离作为相似度，它是求对应某一初始聚类中心向量 $V$ 最优分类，使得评价指标 $J$ 最小。算法采用误差平方和准则函数作为聚类准则函数[18]。

聚类分析的目的在于把集中的数据划分为一系列有意义的子集(或称类)，使得每个子集中的数据尽量“相似”或“接近”，而子集与子集间的数据尽可能有较大差异[19]。聚类分析通常遇到的困难是：

1) 聚类个数 $K$ 并不能预先确定。我们需要找到一个有效性指标，来确定最佳聚类个数。

2) 类中心点不能预先知道，需要用某种方法初设“种子”。

K-means 聚类算法的主要思想可用下式描述：

$$\text{Minimize } J\left(X,U,V\right)=\sum_{i=1}^{k}\sum_{j=1}^{n}u_{ij}\left\|x_j-v_i\right\|^2 \tag{1}$$

其中：$n$ 是给定数据集中的数据个数，$k$ 是聚类个数。

$X=\left\{x_1,x_2,\cdots,x_n\right\}\subset R^S$ 是给定的数据集，$V=\left\{v_1,v_2,\cdots,v_k\right\}\subset R^S$ 是类中心点，用 $X_1,X_2,\cdots,X_k$ 表示 $k$ 个类，$n_i$ 表示 $X_i$ 中数据个数，$U=\left(u_{ij}\right)_{k\times n}$ 是聚类矩阵，由 $x_j$ 和 $X_i$ 的成员关系组成，$d\left(x,y\right)=\left\|x-y\right\|,x,y\in R^S$ 是一个距离函数(例如，欧几里得距离)。为了极小化 $J\left(X,U,V\right)$ 类中心点 $v_i\left(i=1,2,\cdots,k\right)$ 和成员关系矩阵 $U$ 需要用以下的迭代公式逐步计算：

$$u_{ij}=\begin{cases}1;\ \left\|x_j-v_i\right\|\le\left\|x_j-v_h\right\|,h=1,2,\cdots,k,h\ne i\\0;\ \text{否}\square\end{cases} \tag{2}$$

而

$$v_i=\frac{\sum_{j=1}^{n}u_{ij}x_j}{\sum_{j=1}^{n}u_{ij}}=\frac{\sum_{x_j\in X_i}x_j}{n_i} \tag{3}$$

通过某些方法(如随机抽取)初始化类中心点然后通过方程式(2)计算成员关系矩阵 $u_{ij}\left(i=1,2,\cdots,k,j=1,2,\cdots,n\right)$。K-means 算法就是基于(2)、(3)的迭代过程：$V_{t-1}\to U_t\to V_t$，直到 $\left\|v_t-v_{t-1}\right\|\le\varepsilon$。其中 $\varepsilon$ 是给定终止条件。最终的聚类结果通过成员关系矩阵确定，即如果 $u_{ij}=1$，则说明 $x_{ij}\left(j=1,2,\cdots,n\right)$ 属于类 $X_i\left(1\le i\le k\right)$ [20]。

将此算法归纳如下：

算法 A

1) 输入聚类数口 $k$，确定距离函数，给定迭代终止条件 $\varepsilon$；

2) 初始化中心点谓 $v_i^0\left(i=1,2,\cdots,k\right)$；

3) 利用方程式(2)计算 $u_{ij}\left(i=1,2,\cdots,k,\ j=1,2,\cdots,n\right)$；

4) 利用方程式(3)计算新的中心点 $v_i^1\left(i=1,2,\cdots,k\right)$；

5) 如果 $\max\limits_{1\le i\le k}\left\|v_i^0-v_i^1\right\|\le\varepsilon$，那么迭代终止，转向 6)，否则令：$v_i^0=v_i^1\left(i=1,2,\cdots,k\right)$ 转向 3)；

6) 出分类结果：类中心点 $v_i^1\left(i=1,2,\cdots,k\right)$ 和成员关系矩阵 $U$；

7) 终止。

确定最佳聚类数目的算法 B：

1) 选取 $k_{\min}$ 和 $k_{\max}$；

2) For $k = k_{\min}$ to $k_{\max}$；

① 初始化类中心点；

② 利用算法 A，更新计算 $U^k$ 和 $V^k$；

③ 检查终止条件，如不满足，则转向②；

④ 利用方程式(4)计算 $V_{km}(k)$，转向 2)；

3) 选择 $k_{opt}$ 使聚类有效函数 $V_{km}(k)$ 达到最优(最小)；

4) 输出聚类结果：类中心点矽 $V_{k_{opt}}$，成员关系矩阵 $U_{k_{opt}}$，最佳聚类数目 $k_{opt}$；

5) 终止。

### 5.1. 利用 HSV 空间模型进行特征提取

本节主要利用了上文中介绍的 H (色调)、S (饱和度)、V (亮度)的空间模型，在 matlab 中实现了特征提取。我分别选取了 50 张春天、41 张夏天、50 张秋天和 48 张冬天天气图作为特征提取的样本，

通过 matlab 软件的处理得到了各类天气图像的色调平均值、方差(主要为 H 通道数据)，以及每张天气图像的各个像素点的方差(去除异常值后)和各类天气图像的方差的均值。在对时间季节图像的 HSV 特征数据提取后，我对这三类数据又分别做了求平均和标准差的处理，最后，时间季节 H、S、V 三个通道的特征数据体情况如表 1。

由表 1 可知，在时间季节方面的识别-色调是最主要的依据，通过 HSV，我发现在时间季节方面对于秋天、冬天与(春天和夏天)的识别有较好的结果，但是对于春天和夏天的识别明显有较大的误差。

### 5.2. K-Means 聚类法的实现

分类算法是给出一个数据，然后判断这个数据属于已分好的类中哪一类。在 4.1 中，我所做的工作就是基于 HSV 空间模型给时间季节的特征数据提取，并提出一个分类的依据。

聚类算法是给一大堆原始数据，然后通过算法将其中具有相似特征的数据聚为一类。这个过程在我理解看来，也可以是一种特征数据的整合、自动聚类实现特征数据的提取。

本节采用了 2.2 中介绍的 K-means 的聚类算法，在 Matlab 中实现对数据的聚类算法。

由表2可以看出在时间季节的特征提取上聚类算法的效果是比较明显的，它弥补了利用HSV识别中春天和夏天特征分辨不清楚的缺陷。

### 5.3. K-NN 最邻近法的实现

在实现了特征提取，现在的问题就是如何对处理后的特征数据进行分类，本节本节利用了第二章中介绍的 K-邻近分类算法.

我选取了 50 张春天、50 张夏天、50 张秋天和 50 张冬天图像作为待识别图像样本库，经过对每一张图像的识别分类，最终各项识别准确度如表 3。

## 6. 总结

不同的时间季节有不同的特征，根据季节的特征，我们在计算机上利用各种方法可以清楚的识别不同的季节。本文利用 HSV 图像色彩模型，主要对各个季节的图片的特征进行分析和研究，进行颜色间的对比，比较其色彩值的平均值与方差，运用光学图像处理技术，采用近邻分类的方法，对比各季节图像

**Table 1.** Time-season picture HSV feature data
**表 1.** 时间季节图像 HSV 特征数据

| 序号 | 类别 | H_avg | H_var | S_avg | S_var | V_avg | V_var |
|---|---|---|---|---|---|---|---|
| **1** | 春天 | 0.2716 | 0.0944 | 0.5652 | 0.2214 | 0.4933 | 0.2123 |
| **2** | 夏天 | 0.3590 | 0.0923 | 0.5568 | 0.2146 | 0.4317 | 0.2074 |
| **3** | 秋天 | 0.1559 | 0.1473 | 0.6831 | 0.2164 | 0.5577 | 0.2362 |
| **3** | 冬天 | 0.5584 | 0.1139 | 0.2593 | 0.1574 | 0.6526 | 0.2088 |

注：H_avg 表示色调均值；H_var 表示色调标准差；S_avg 表示饱和度均值；S_var 表示饱和度标准差；V_avg 表示亮度均值；V_var 表示亮度标准差。

**Table 2.** The seasonal tonal feature data of clustering algorithm
**表 2.** 聚类算法的季节色调特征数据情况

| 序号 | 类别 | 色调均值 | 色调标准差 |
|---|---|---|---|
| **1** | 春天 | 0.2645 | 0.0856 |
| **2** | 夏天 | 0.3610 | 0.0827 |
| **3** | 秋天 | 0.1721 | 0.1140 |
| **4** | 冬天 | 0.5741 | 0.0864 |

**Table 3.** Image recognition accuracy
**表 3.** 图像识别准确度

| 天气类别 | 样本数 | 识别准确数 | 准确率 |
|---|---|---|---|
| 春天 | 50 | 40 | 80% |
| 夏天 | 50 | 38 | 76% |
| 秋天 | 50 | 44 | 88% |
| 冬天 | 50 | 45 | 90% |

的特征差异，对所选图像进行分类与识别，进而达到自动识别季节的目的。实验结果证实了此研究方法的可行性，在季节识别上能够达到较高自动识别图片的目标。

## 致　谢

论文得到国家自然科学基金(No.61471412, 61771020)的支持。

## 参考文献 (References)

[1] Worf, W. (1996) Key Frame Selection by Motion Analysis. 1996 *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Atlanta, 9 May 1996, 1228-1231.

[2] Zhang, H.J., Wu, J.H., Zhang, D., *et al.* (1997) An Integrated System for Content-based Video Retrieval and Browsing. *Pattern Recognition*, **30**, 643-658. https://doi.org/10.1016/S0031-3203(96)00109-4

[3] Han, J., Kamber, M. and Pei, J. (2006) Data mining: Concepts and Techniques. Morgan Kaufmann, Burlington, Massachusetts.

[4] 滕敏, 卫文学, 滕宁. K-最近邻分类算法应用研究[J]. 软件导刊, 2015(3): 44-46.

[5] [美]冈萨雷斯, [美]伍兹. 数字图像处理[M]. 第 3 版. 阮秋琦, 译. 北京: 电子工业出版社, 2001.

[6] 张彩华. 基于颜色和纹理特征的图像检索[D]: [硕士学位论文]. 哈尔滨: 哈尔滨理工大学, 2010.

[7] 顾宁. 基于颜色和纹理特征的图像检索算法研究[D]: [硕士学位论文]. 南京: 南京邮电大学, 2011.

[8] 葛静祥. 图像纹理特征提取及分类算法研究[D]: [硕士学位论文]. 天津: 天津大学, 2010.
[9] 王惠明, 史萍. 图像纹理特征的提取方法[J]. 中国传媒大学学报(自然科学版), 2006, 13(1): 49-52.
[10] 幸锐. 基于纹理的图像聚类研究[D]: [博士学位论文]. 杭州: 浙江大学, 2009.
[11] 谢菲. 图像纹理特征的提取和图像分类系统研究及实现[D]: [硕士学位论文]. 成都: 电子科技大学, 2009.
[12] 蒋良孝. 朴素贝叶斯分类器及其改进算法研究[D]: [博士学位论文]. 武汉: 中国地质大学, 2009.
[13] 沈国杰. 一种基于模糊关联分类的遥感图像分类方法[J]. 计算机研究与发展, 2012, 49(7): 1500-1506.
[14] 肖靓. 基于支持向量机的图像分类研究[D]: [硕士学位论文]. 上海: 同济大学, 2006.
[15] 云峰, 周玲, 于俊清, 徐涛, 管涛. 基于局部特征聚合的图像索引方法[J]. 计算机学报, 34(11): 2224-2233.
[16] 陈慧. 基于内容的图像检索技术研究[J]. 福建金融管理干部学院学报, 2007(6): 55-59.
[17] 刘燕, 邝颖杰. 基于混合索引的图像检索系统的设计与实现[J]. 农业网络信息, 2007(6): 34-36.
[18] 魏志静. 基于人工神经网络的分类方法研究及其在个人信用评估中的应用[D]: [硕士学位论文]. 山东师范大学, 2007.
[19] 孙秀亮. 基于属性加权的选择性朴素贝叶斯分类研究[D]: [硕士学位论文]. 哈尔滨: 哈尔滨工程大学, 2013.
[20] 王丹, 吴孟达. 粗糙模糊 C-均值算法及其在图像聚类中的应用[J]. 国防科技大学学报, 2007, 29(2): 76-80.