改进 FCM 的图像聚类方法①

周俊祥

(商丘师范学院 计算机科学系，商丘 476000)

摘 要：对传统 FCM 算法的隶属度函数进行了改进，改进后的算法有效降低了孤立点对图像数据聚类结果的影响。通过灰度-梯度共生矩阵对图像进行纹理特征提取，利用主分量分析法对提取后的图像高维特征进行降维处理，结合本文改进的 FCM 图像聚类算法对预处理后的图像数据进行聚类。实验证明，该方法具有较好的聚类效果，且能以较少的迭代次数达到全局最优。

关键词：FCM 算法；图像聚类；隶属度函数；主分量分析法

Image Clustering Based on Improved FCM Algorithm

ZHOU Jun-Xiang

(Department of Computer Science and Technology, Normal University of Shangqiu, Shangqiu 47600, China)

Abstract: In this paper, the traditional FCM algorithm membership function was improved. The improved algorithm can reduce the isolation point of the image data clustering results. In this paper, Gray-gradient co-occurrence matrix of the image texture feature extraction using principal component analysis on the extracted high-dimensional feature image to reduce the dimensions, combined with this improved FCM clustering algorithm to the image after the image data preprocessing clustering. Experiments show that the method has better clustering results, with fewer iterations and can reach the global optimum.

Key words: FCM algorithm; image clustering; membership function; principal component analysis

1 引言

作为一种有效的数据分析方法，聚类算法已经被广泛应用于图像处理、客户关系管理、市场营销、医学、模式识别等领域[1]。聚类算法在执行过程中不能获得任何关于预先定义的数据项的类属信息，因而通常被看作是一种无监督学习方法。模糊聚类算法实际上是硬聚类和模糊集合理论相结合的延伸和推广。其中，模糊 C-均值聚类 FCM 算法(Fuzzy C-Means Clustering)是应用最广的聚类算法，该算法是基于划分的方法，目的是把数据分到不同的类或簇，在划分好的结果中同一簇中的对象之间相似性很大，而不同簇间的对象相异性很大。FCM 算法是普通 C-均值算法的改进，普通 C 均值算法对于数据的划分是硬性的，模糊聚类是柔性的划分，该聚类通过使用隶属函数，使得可以把每一个对象分配给所有的聚类，模糊聚类的结果是每个对象最终可能属于多个聚类，每个对像对每个聚类分配一个隶属度。但是，传统的 FCM 算法存在对孤立点比较敏感的缺陷，本文对传统 FCM 算法的隶属度函数进行了改进，改进后的算法有效弥补了该缺陷。

随着社会信息化程度的不断加深，图像处理相关技术应用更加广泛，

随着社会信息化程度的不断加深，图像处理相关技术应用更加广泛，而图像聚类是图像处理技术和模式识别中的一个重要环节。目前，该技术已经渗透到很多领域：天文学中的望远镜图像聚类和卫星遥感图像聚类；医学中的心电图聚类和医学图像聚类；军事领域中的航空摄像的聚类等。本文提出了一种改进的

① 基金项目:国家高技术研究发展计划(863)(2009AA01Z302)
收稿时间:2010-05-01;收到修改稿时间:2010-05-26

FCM 算法对图像数据进行聚类。实验证明，该聚类方法具有速度快、聚类效果好的特点。文章的思路是：(1)对聚类的图像提取图像特征；(2)对所提取的特征进行特征选择；(3)基于改进的 FCM 算法的图像聚类；(4)实验；(5)结论。

1 图像纹理特征提取与灰度共生矩阵

1.1 图像纹理特征提取

图像特征提取是计算机视觉和图像处理中的一个概念，它指的是使用计算机提取图像信息，决定每个图像的点是否属于一个图像特征，特征提取的结果是把图像上的点分为不同的子集，这些子集通常属于连续的曲线、连续的区域或者是孤立的点。

纹理是景物的一个重要特征。图像的纹理是指在图像上表现为灰度或颜色分布的某种规律性，这种规律性在不同类别的纹理中有不同特点[2]。通常图像的纹理特征可以分为准规则纹理和规则纹理两大类。其中，准规则纹理是某种灰度或颜色的分布，该分布在空间位置上的反复出现形成纹理，这类纹理的纹理基元没有明确的形状，存在着局部不规则和整体规律性的特点，又被称为自然纹理。规则纹理是由明确的纹理基元经有规则排列而成，又被称为人工纹理。可以根据纹理类别的不同而采用不同纹理特征提取的方法，提取方法有结构分析方法和统计分析方法两种。本文是基于图像的纹理特征对图像特征进行提取，采用了统计分析方法研究像素的二阶统计特性，即对于输入的图像进行纹理分析，用灰度-梯度共生矩阵来提取图像的纹理特征。

1.2 灰度-梯度共生矩阵

图像的纹理特征信息既可以是灰度本身的信息[3]，又可以是灰度变化的梯度信息。灰度-梯度共生矩阵元素 H(x,y)定义为正规化的灰度图像 F(m,n)和正规化的梯度图像 G(m,n)中共同具有灰度值为 x 和梯度值为 y 的总像点数。灰度是构成图像的基础，而梯度是构成图像轮廓的要素。灰度-梯度共生矩阵是用灰度和梯度的综合信息提取纹理特征，它考虑像素灰度与边缘梯度的联合统计分布。因此，灰度-梯度共生矩阵反映了图像的灰度和梯度的分布规律,且表现了各像素点与其邻域像素点之间的空间关系。

设一幅 $M \times M$ 灰度图像 $\{f(m,n); m=0,1,...,M-1, n=0,1,...,M-1\}$ 。为了简化计算，进行灰度进行正规化处理为：$F(m,n)=INT(f(m,n)\times L_f / f_{\max})+1$，其中，INT()为取整运算函数，$L_f$ 为规定的灰度级数，$f_{\max}$ 为图像 $f(m,n)$ 的最大灰度。同理，将梯度图像 $\{g(m,n); m=0,1,...,M-1, n=0,1,...,M-1\}$ 进行正规化处理为 $G(m,n)=INT(g(m,n)\times L_g / g_{\max})+1$，其中，$L_g$ 为规定的灰度级数目，$g_{\max}$ 为梯度图像 $g(m,n)$ 的最大梯度。因此，在正规化后的基础上，灰度-梯度共生矩阵 $\{H(x,y); x=0,1,...,L_f-1, y=0,1,...,L_g-1\}$，H(x,y)定义为集合：

$\{(x,y)\,|\,G(m,n)=x, F(m,n)=y; i,j=0,1,...,M-1\}$

中的元素数目，即灰度为 x，梯度为 y 的总像素个数。最后对灰度—梯度共生矩阵进行归一化处理，使其各元素之和为 1。

利用灰度-梯度共生矩阵我们计算了 15 个二次统计特征参数[4]：小梯度优势、大梯度优势、灰度分布不均匀性、梯度分布不均匀性、能量、灰度平均、梯度平均、灰度方差、梯度方差、相关、梯度熵，灰度熵、混合熵、惯性、以及逆矩差。因此，我可以得到 15 个纹理特征集合。

2 图像纹理特征选择

纹理特征选择的主要目的是降维。其主要思想是将原始样本投影到一个低维特征空间，得到最能反应样本本质或进行样本区分的低维样本特征。使用灰度-梯度共生矩阵对图像进行的纹理提取出来的纹理特征，其维数大，有可能存在一些与聚类目标无关冗余信息，从而导致聚类效果降低，增大系统的开销。因此，为了消除其冗余信息，降低特征维数之间的相关性，需要对纹理特征进行选择。本文使用主分量分析法对其进行选择，得到最优的纹理特征子集。

主成分分析(PCA)也称为主分量分析[5]，该方法是使用降维的思想，在损失很少信息的前提下把多指标转化为少数几个综合指标的多元统计分析方法，把转化后的综合指标称为主成分，每个主成分互不相关且是原始变量的线性组合。设输入的原始数据 x 为 m 维,通过线性变换 y =Wx 投影到一个低维特征空间,得到 n 维数据 y(n<m)。通过寻找线性变换矩阵 W，该方法使降维后产生的误差在最小均方误差下最优,即使得主分量部分具有较大的能量。因此，在不丢失信息的前提下，主成分分析法可以使复杂的问题得到简化，能

进一步很好的聚类。

为了消除数据量纲和数量记得影响，需要输入 m 个样品且每个样品有 q 项指标的原始数据进行标准化变换，生成标准化样本数据矩阵 $Y_{m\times q}$；计算标准化数据矩阵每两维特征值间的相关系数，得到相关系数矩阵 $R_{q\times q}$；使用雅可比方法求解相关系数矩阵的特征方程 $|\sigma I - R| = 0$，求出特征根并对其排序，得到 $\sigma_q \le \sigma_{q-1} \le ... \le \sigma_2 \le \sigma_1$，然后分别求出每个特征根相对应的特征向量，得到特征向量矩阵 $W\left(W = (W_{ij})_{q\times q}\right)$，在不丢失太多原始数据信息的基础上，按累计方差的贡献率的大小确定取前 n 个主成分，从而实现降维的目的。计算前 n 个主成分的样本值，确定新的矩阵 $Y_{m\times n}$，即降维后的样本数据为 $Y_{ij} = \sum_{k=1}^{q} y_{ik} w_{kj}$ (i=1,2⋯.m,j=1,2⋯n)，用其进行聚类，从而简化问题的复杂度。

3　基于改进的FCM的图像模糊聚类

FCM 模糊聚类算法是一种基于目标函数的动态优化算法，它是一种能自动对数据样本进行聚类的方法[6,7]。根据图像像素和聚类中心加权相似性测度，该方法对目标函数进行迭代聚类。聚类结果使得被划分到同一类中的对象之间相似度最大，而不同类对象之间的相似度最小。

使用改进的 FCM 算法对选择后得到的最优特征子集进行聚类，优化了算法的速度和图像聚类的精确度。由于传统的 FCM 算法对孤立点比较敏感，所以为了降低孤立点对聚类结果的影响，本文对数据隶属度增加一个权值，使隶属度值小的数据对象降低他们对聚类中心的影响，隶属度值高的数据对象增大他们对聚类中心位置的影响。隶属度的改进函数如式(1)：

$$M_{ij} = \lambda U_{ij} + (1-\lambda)U_{ij}^2 \qquad (1)$$

式中，隶属度函数 $M_{ij} = f(U_{ij})$ 是凹函数。$0 \le \lambda \le 1$，当 $\lambda = 1$ 时，$M_{ij} = U_{ij}$；当隶属度 $U_{ij} = 0$ 时，$M_{ij} = 0$。在[0,1]区间内的改进之后的隶属度值比原来有一定的减少。算法的迭代过程中，隶属度值越小，对应的 M_{ij} 减少的越多。利用聚类中心公式之后，隶属度小的图像数据对聚类中心的影响就降低了。隶属度 M_{ij} 值的调节幅度随着 λ 的增大而减小。且对 M_{ij} 调节幅度越小，聚类结果的精度就越精确，在本算法中，取参数 λ 为 0.4。

FCM 聚类算法的思路为：设数据集 X 中含有 n 个待聚类的样本，表示为 x_k (k=1,2,⋯,n)。将数据集 $X = (x_1, x_2, ..., x_n)$ 划分为 c 类，X 中的任意样本 x_k (k=1,2, ⋯ ,n) 对第 i(i=1,2, ⋯ ,c) 类的隶属度 $\mu_{ik}(0 \le \mu_{ik} \le 1)$，则该分类结果可以用一个 $c \times n$ 阶的矩阵 U 来表示，该矩阵称为模糊聚类矩阵：

$$\mu_{ik} \in [0,1], (k = 1,2,...,n; i = 1,2,...,c)$$

$$\sum_{k=1}^{c} \mu_{ik} = 1, (\forall k, k = 1,2,...,n; i = 1,2,...,c)$$

$$0 < \sum_{i=1}^{n} \mu_{ik} < n, (\forall i, k = 1,2,...,n; i = 1,2,...,c)$$

FCM 聚类算法的目标函数：

$$\min J_m(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{c} (\mu_{ik})^m (d_{ik})^2$$

式中，$U = [\mu_{ik}](k = 1,2,...,n; i = 1,2,...c)$ 为模糊聚类矩阵；$V = (v_1, v_2, ..., v_c)$ 为 c 个聚类中心的集合；$m \in (1, \infty)$ 为加权指数，$d_{ik} = \|x_k - v_i\|$ 为第 k 个样本点 x_k 到第 i 个聚类中心的欧氏距离。

改进的 FCM 算法如下：

1) 初始化：设定聚类类别数 c，加权指数 m，设定迭代停止阈值 $\varepsilon > 0$，算法的最大迭代次数 $T_{\max}$，初始化聚类中心 $V^{(1)}$，并令 k=1；

2) 根据公式 $\mu_{ik} = \dfrac{1}{\sum_{j=1}^{c}\left(\dfrac{d_{ik}}{d_{jk}}\right)^{\frac{2}{m-1}}}$ 由 $V^{(k)}$ 计算 $U^{(k)}$；

3) 改进隶属度矩阵 U

使用式 $M_{ij} = \lambda U_{ij} + (1-\lambda)U_{ij}^2$ 对隶属度函数改进，降低孤立点对聚类中心的影响，其中参数 λ=0.4。

根据公式 $v_i = \dfrac{\sum_{k=1}^{n}(M_{ik})^m x_k}{\sum_{k=1}^{n}(M_{ik})^m}$ 由计算；

4) 如果 $\|V^{(k+1)} - V^{(k)}\| < \varepsilon$，停止；否则，k=k+1，转到步骤 1)。

通过 2)和 3)反复修改隶属度和聚类中心，当算法收敛时理论上就得到了各类的聚类中心，以及各个样本对于各模式类的隶属度从而完成了模糊聚类划分。

4 实验

本文实验是对卫星遥感图进行聚类，其中，实验一是从某市卫星遥感图像库中选出 72 幅图像作为实验数据，其中，水体 22，建筑物 30 幅，树木 20。 根据本文的算法对图像数据进行聚类，每一类放在一起，实验结果如图 2 所示：

图 2 基于 FCM 的图像聚类

由图 2 可知，图的上方为树木分类，正确数为 18 幅；中间为建筑物分类，正确数为 24 幅；下方为水体分类，正确数为 20 幅。总体分类精度是 86.1%。

实验二是对该市整体遥感图进行聚类，其中，红色代表该市的建筑物，绿色代表树木，蓝色代表水体，聚类后的图像与原图像如图 3 所示：

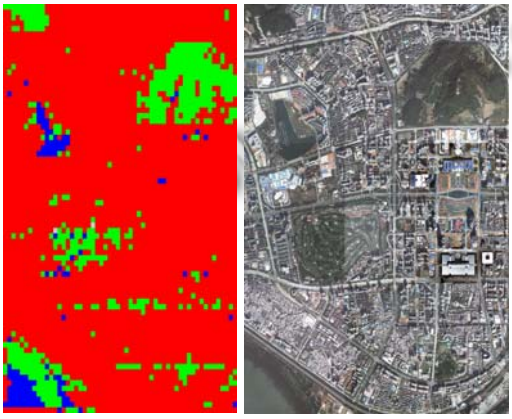

图 3 RGB 聚类显示图与原始图比较

由图 3 可知，使用改进的 FCM 聚类算法得到的 RGB 聚类显示图也具有较高的分类精度。

5 结论

本文针对传统的 FCM 对孤立点敏感的缺点，对 FCM 的隶属度函数进行了改进。首先，使用灰度—梯度共生矩阵法对图像进行了纹理特征提取，得到每幅图像的 15 维特征向量。其次，采用 PCA 主向量分析法对特征向量进行特征选择，即对特征向量降维。最后，结合改进的 FCM 算法对降维后的图像数据进行聚类。本文是以卫星遥感图像作为实验数据对图像进行聚类。实验结果表明本文的方法具有较好的聚类效果和较高的聚类效率。

参考文献

1 Lim YW, Lee SU. On the Color Image Segmentation Algorithm Based on the Thresholding and the Fuzzy c-means Techniques. Pattern Recognition, 1990,23(9):935−951.

2 胡召玲,李海权,杜培军.SAR 图像纹理特征提取与分类研究.中国矿业大学学报,2009,38(3):422−427.

3 唐玮,朱华,王勇.分形和空间灰度共生矩阵联合评价断口形貌研究.中国矿业大学学报, 2006,35(4):530−534.

4 窦唯,刘占生.基于灰度-梯度共生矩阵的旋转机械振动时频图形识别方法.振动工程学报,2009,22(1):85−91.

5 智晶,张冬梅,姜鹏飞.基于主成分的遗传神经网络股票指数预测研究.计算机工程与应用,2009,45(26):210−212.

6 刘志勇,耿新青.基于模糊聚类的文本挖掘算法.计算机工程,2009,35(5):44−49.

7 陈新泉.特征加权的模糊 C 聚类算法.计算机工程与设计,2007,28(22):5329−5333.