文章编号: 1001- 2486(2008) 06- 0073- 05

基于半监督 FCM 聚类算法的卫星云图分类*

来 旭, 李国辉, 张 军
(国防科技大学 信息系统与管理学院, 湖南 长沙 410073)

摘 要: 针对卫星云图的特点在分类特征集中采用了一种新的特征——差异化特征, 该特征反映了云图的内部结构特点, 并且具有良好的鲁棒性, 能有效地避免云团位置变化对特征的影响。将半监督思想引入到模糊 C 均值聚类方法(FCM), 克服了单纯的 FCM 方法未考虑领域知识导致的聚类结果的盲目性。半监督 FCM 方法在聚类过程中加入少量的由领域专家标记的样本, 引入专家的领域知识, 通过与这些带有类标记的样本进行相似性比较, 引导 FCM 方法的聚类过程。试验结果表明, 基于具有差异化特征的云图特征集, 半监督 FCM 方法能有效地提高云图分类的准确率。

关键词: 差异化特征; 半监督 FCM; 卫星云图分类
中图分类号: TP391 文献标识码: A

Satellite Cloud Images Classification Based on Semi-supervised FCM Method

LAI Xu, LI Gue-hui, ZHANG Jun
(College of Information System and Management, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: The paper proposes a new classification feature: DI(Diversity Index), considering the characteristics of satellite cloud image. The DI feature presents the structure of cloud effectively and it has a good robustness. The DI feature avoids the influence exerted by the variety of cloud positions. This paper proposes the semi-supervised FCM (SSFCM) method in the domain of satellite cloud images classification. The SSFCM method overcomes the blindness brought by the FCM method without considering the domain knowledge. The SSFCM method uses a small number of samples labeled by experts to direct the clustering process through comparing with the labeled samples in terms of similarity. These labeled samples represent the domain knowledge. The experiments demonstrate that the SSFCM method improves the accuracy of cloud classification based on the DI feature.

Key words: diversity index (DI); semi-supervised FCM; satellite cloud classification

卫星遥感技术是目前研究大气云团时空属性最有效的手段。云图展示了大气中各种云类的形态, 蕴含着丰富的天气演变信息, 实时反映了大气中正在进行的动力和热力过程。能够准确识别云图中云团的种类及其分布, 对提高灾害性天气的监测、辅助防洪防汛的决策等都有着重要的现实意义。过去, 主要依靠专家的专业知识和经验判读云图中云团的种类。随着计算机存储技术、图像处理以及人工智能技术的发展, 自动分析识别云类日益成为研究的热点。建立客观准确的云分类模型, 使用恰当的分类方法是实现云自动分类的前提和基础。早期 Koffer 等[1] 使用简单的阈值法区分云和陆地, 后来 Desbois 等[2] 提出了光谱特征空间的概念, 创建了盒式分类法, 云分类在简单阈值分类基础上得到很大进展。Welch 等[3] 在运用云的纹理特征进行云分类方面做了大量工作。国内的郁凡等[4] 在前人工作的基础上, 提出了单位隶属特征空间的概念, 利用单位特征空间对各云类分布区域进行划分和拟合, 进一步提高云自动分类的准确性。

云的成因和聚散生消非常复杂, 除一些特征明显的典型云类外, 还存在云类之间的过渡区域和处于生长、消亡阶段等特征较为模糊的云系, 它们在卫星云图上表现出的灰度和纹理特征比较复杂, 硬性把

* 收稿日期: 2008- 02- 18
基金项目: 国家自然基金资助项目(60473116)
作者简介: 来旭(1979—), 男, 博士生。

这些云归为某一类别并不合理,易产生误判。常用的统计分类方法在处理这类特征不很明显的问题时表现出较大的局限性。本文采用模糊C均值聚类方法(FCM),通过引入隶属度的概念,克服硬性云分类的不足。FCM是一种非监督的分类方法,分类过程中不需要人为的干预,但该方法也存在不足,由于分类的过程完全基于数据自身的相似性,并未结合领域知识,有可能获得无效的分类结果。针对FCM的不足之处,本文提出了半监督的FCM算法,它在经典的FCM算法的基础上,加入部分先验信息,改进后的算法更适合处理卫星云图。

1 云图特征空间的确定

在云分类的过程中,特征空间选定的有效性将直接影响分类的准确性。随着图像处理技术的日趋成熟,纹理特征被引入到云图的识别中并体现出重要的作用。Haralick等[5]于1973年提出灰度共生矩阵统计量,基于灰度共生矩阵提取了包括能量、熵、惯量等在内的28个纹理特征。随着卫星遥感技术发展,云图分辨率不断增强,纹理特征的使用越来越普遍。但如何从丰富的遥感信息中提取有效特征,采用何种特征能有助于云分类准确性的提高,这些都是人们关注的问题。

专家在用肉眼判别云种类时主要是以云的光滑程度、明暗程度等信息作为判断依据。云的种类、厚度不一,使得云顶表面有的表现为光滑,有的表现为多起伏的斑点和皱纹,有的表现为纤维状。本文采用了两种常用的基于灰度共生矩阵的纹理特征:标准差和熵,还采用了另一种较新的特征:DI(Diversity Index),该特征曾被用于生态系统内在差异性的研究[6],本文将它用于描述不同云类内部结构的差异。DI特征的理论基础是信息学中经典的香农公式:

$$H = -\sum_{i}(p_i \cdot \log_2 p_i) \tag{1}$$

式中,p_i 表示对象处于第 i 种状态的概率。在实际的运用中,p_i 常采用(2)式计算获得,N 表示样本的总数,N_i 表示处于第 i 种状态的样本数,假设共有 S 种状态,则 $\sum_{i=1}^{S} N_i = N$。

$$p_i = f_i = N_i / N \tag{2}$$

将(2)式代入(1)式后,计算公式变为:

$$H' = -\sum_{i}^{S}(f_i \cdot \log_2 f_i) \tag{3}$$

以上述公式作为理论基础,通过下列步骤的处理,将从云图样本中获取DI特征:

步骤1 首先按照8×8大小的尺寸对原始云图进行分块处理。原始云图是128×128的大小,分块处理后获得256个图像块,因此(2)式中 N 的取值为256。

步骤2 计算每个图像块的标准偏差 σ,找出所有图像块中 σ_{max} 和 σ_{min},等间距地将[σ_{max},σ_{min}]区间划分为10个子区段,因此公式(3)中 S 的取值为10。

步骤3 将每个图像块计算得到的 σ_i 映射到10个子区段上,以区段标号替代 σ_i,如图1所示。

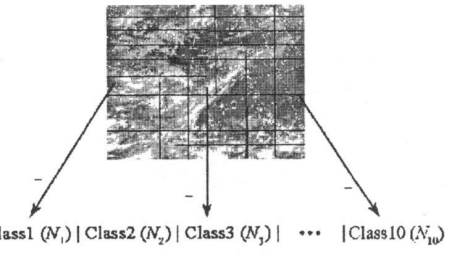

图1 图像块标准偏差投影

Fig.1 Mapping of standard deviation of image block

步骤4 统计每个区段内图像块个数,代入公式(2)、(3)中,计算DI的值。

$$f_i = N_i/256, \quad \mathrm{DI} = -\sum_{i}^{S=10}(f_i \cdot \log_2 f_i)$$

DI特征相对于其他云图特征具有自身的优势:(1)具有良好的鲁棒性。云具有流动性,同一种云的特征值会随云位置的改变而变化,最终被误判为其他类别的云。DI特征的计算过程可有效避免云位置变化的影响。(2)本文采用的样本尺寸较大,常规纹理特征的计算方法中都存在平均化过程,如果直接将其用于云图样本将会抹去很多云内的细节信息。DI特征的处理过程可保留更多的云的结构细节信

息。将云分类时应选取充足的特征, 必须对特征进行分析选择, 形成云分类有效的识别特征。通过分析云图样本投影在特征空间内的散点图, 确定对云类具有良好区分能力的特征空间。图 2 反映各类云的样本在特征空间的分布情况, 圆形区域反映的是不同种类云的样本混杂的情况。“熵 —*DI*”特征空间混杂程度最轻微, 因此本文确定在“熵 —*DI*”特征空间完成云的分类实验。

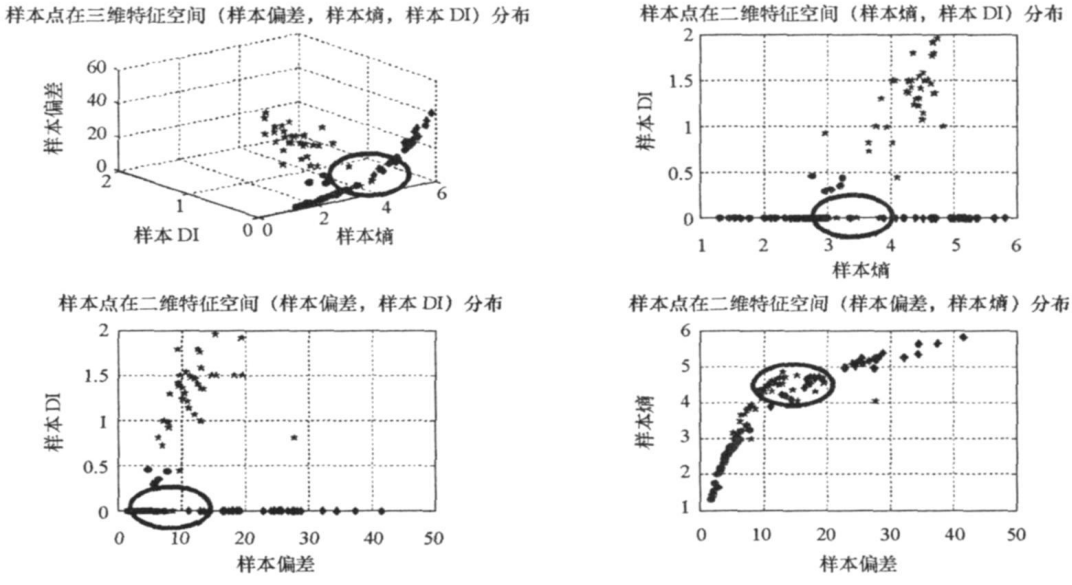

图 2　样本点在特征空间内的分布

Fig. 2　Distribution of samples in the feature space

2　基于特征空间的半监督 **FCM** 云图聚类研究

2.1　模糊 C 均值聚类(FCM) 原理介绍

常规 *FCM* 算法是一种无监督的聚类方法, 它通过最小化目标函数来实现数据聚类, 其目标函数如下[7]:

$$J_m(\boldsymbol{U}, \boldsymbol{V}) = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}^{m} d_{ik}^{2}, \quad m \geqslant 1 \tag{4}$$

其中, $\boldsymbol{X}$ 是图像像元集合, n 是像元个数; $\boldsymbol{V}$ 是聚类中心集合, c 是类别数; $\boldsymbol{U}$ 是隶属度矩阵, u_{ij} 代表图像像元 x_j 对于类别 v_i 的隶属度; m 是加权指数, 它控制类别之间的分享程度, m 越大, 所得的分类矩阵模糊程度就越大。$d_{ij} = \|x_j - v_i\|$是样本点 x_j 和聚类中心 v_i 的欧氏距离。

步骤 1　确定聚类数 c, 加权指数 m, 阈值 ε, 最大迭代次数;

步骤 2　确定样本的初始隶属度矩阵 $\boldsymbol{U}$;

步骤 3　依次取迭代步数 $b = 0, 1, 2, \cdots$, 根据公式(5), 对 u_{ij}, v_i 进行反复迭代:

$$u_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{kj}}\right)^{\frac{2}{m-1}}}, \quad v_i = \frac{\sum_{j=1}^{n} u_{ij}^{m} x_j}{\sum_{j=1}^{n} u_{ij}^{m}} \tag{5}$$

步骤 4　当 $\max_{ij}\|u_{ij} - \hat{u}_{ij}\| < \varepsilon$或者迭代次数大于最大迭代次数时, 迭代结束。

2.2　改进的半监督 **FCM** 云图聚类方法

经典的FCM 算法中, $d_{ij} = \|x_j - v_i\|$为欧氏距离, 适用于各向同性或球体分布。但图 2 反映的云类样本在特征空间的分布不符合各向同性或球体分布, 因此本文采用基于标准协方差矩阵的Mahalanobis 距离[8]替代欧式距离用于FCM 算法。

$$d_{ij}^2 = (\boldsymbol{x}_j - \boldsymbol{v}_i)^{\mathrm{T}} \boldsymbol{C}^{-1} (\boldsymbol{x}_j - \boldsymbol{v}_i) \tag{6}$$

其中, $\boldsymbol{C}$ 为标准协方差矩阵, 如果当矩阵 $\boldsymbol{C}^{-1} = \boldsymbol{I}$ (单位矩阵) 时, 则公式(6)变为欧氏距离。Mahalanobis距离的实质是在计算样本 $\boldsymbol{X}$ 和中心 $\boldsymbol{V}$ 的距离时引入了加权矩阵 $\boldsymbol{C}^{-1}$, 此加权矩阵是经过归一化处理的模糊类内离散度矩阵的逆, 直观来说, 即是在离散度比较大的方向上加权比较小, 而在离散度比较小的方向上加权比较大, 从而可以实现超椭球模糊聚类, 这样能更有效地检测超椭球体分布的各类别。

由于云的情况比较复杂, FCM这类无监督聚类算法的结果难达到较高的准确率。如果在聚类过程中, 某些样本是已知类别的, 那么利用已知类别的样本来影响聚类, 就可以提高聚类的效果, 该过程称为半监督聚类[9]。它与基于监督的分类是不同的, 虽然它们都要加入训练样本, 但分类是根据提供的带有类标号的样本, 通过选择特征参数, 建立判别函数, 然后把图像中各个特征点划归到给定的类中, 而半监督聚类不需要建立判别函数。根据半监督聚类的思想, 本文在使用FCM算法进行聚类的过程中, 可以通过使用一个辅助变量来加入先验信息以影响聚类[10-11], 其迭代过程和FCM算法相同。半监督的FCM算法目标函数如下:

$$J_m(\boldsymbol{U}, \boldsymbol{V}) = \sum_{j=1}^{n}\sum_{i=1}^{c} u_{ij}^m d_{ij}^2 + \alpha \sum_{j=1}^{n}\sum_{i=1}^{c} (u_{ij} - f_{ij} b_j)^m d_{ij}^2 \tag{7}$$

其中, α 是表示监督和无监督程度的参数, 它用未标号和标号的样本数的比值来表示, 在监督与无监督之间维持平衡。b_j 是布尔型变量, 用它来标记标签的和未标签的样本。如果 b_j 为0, 则表示样本 $\boldsymbol{x}_j$ 是未标签的; 如果 b_j 为1, 则表示样本 $\boldsymbol{x}_j$ 是已标签的; 因此已标签样本的隶属度用一个矩阵形式表示: $\boldsymbol{F} = [f_{ij}]$。取 $m = 2$, 半监督FCM方法的迭代公式为:

$$u_{ij} = \frac{1}{1+\alpha}\left[\frac{1+\left(1 - b_j\sum_{k=1}^{c} f_{kj}\right)}{\sum_{k=1}^{c}\frac{d_{ij}^2}{d_{kj}^2}} + \alpha f_{ij} b_j\right], \quad v_i = \frac{\sum_{j=1}^{n} u_{ij}^2 x_j}{\sum_{j=1}^{n} u_{ij}^2} \tag{8}$$

加入部分先验信息之后, 在聚类的过程中, 通过与已知类别的样本进行相似性比较, 提高聚类的准确度。

3 实验结果

本文采用"风云"2号静止气象卫星红外通道拍摄的云图作为研究对象, 选取时段为2007年7月1日00时到2007年9月1日23时内的108幅云图样本, 各类云图的样本情况如表1所示。

表1 各类云图样本数量

Tab. 1 Amount of the cloud sample

云图类别	样本数量
层云	36
积雨云	41
卷云	31

实验主要研究层云、积雨云和卷云这三类云图的自动分类方法, 所以聚类个数 c 为3。实验中选取终止误差 $\varepsilon = 0.001$, 最大迭代次数为50。目前 m 的取值主要是依据经验确定[12], 通常为2。α 是半监督思想的一个重要体现, 它的取值反映了半监督的程度, 当 $\alpha = 0$ 时, 算法退化为常规的FCM聚类; 当 $\alpha = 1$ 时, 聚类变为有监督的分类; 当 $0 < \alpha < 1$ 时, 公式(7)体现了部分带有类表号的样本指导聚类过程的半监督思想。实验分别设定 α 的值为0, 0.1, 0.2, 0.3, 0.4, 0.5, 计算相应的分类准确率。图3描述了 α 和分类准确率之间的关系, 随着 α 的增加, 分类准确率不断提高, α 的值从0.3开始, 分类准确率的变化趋缓。α 的值越大则需要的带有类表号的样本数越大, 实际应用既要保证较高的分类准确率, 同时又要降低手工标记样本标号的工作量, 所以本文最终确定 α 的取值为0.3。

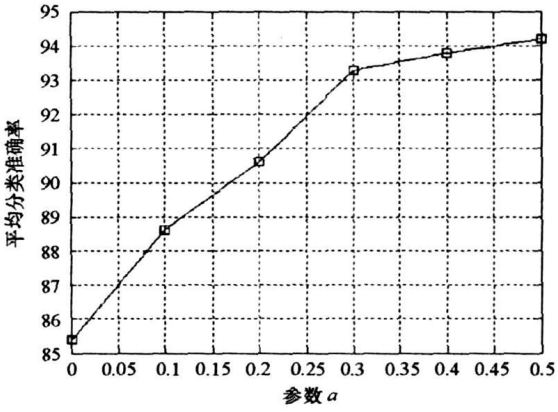

图3 α 对分类准确率的影响

Fig. 3 Effect of the α on the classification accuracy

本文将改进的半监督 FCM 方法所得的聚类结果同 C 均值、FCM 两种聚类方法所得结果进行比较, 如表 2 所示。结果表明半监督 FCM 算法平均聚类准确率优于另两种聚类算法。表 3~ 5 分别对应每种聚类算法结果的混乱矩阵。

表 2　聚类准确率对比表

Tab. 2　Accuracy of different clustering methods

聚类方法	层云正确率	积雨云正确率	卷云正确率	平均正确率
C 均值	100%	73. 2%	61. 3%	78. 2%
FCM	100%	75. 6%	80. 6%	85. 4%
半监督 FCM	94. 4%	85. 4%	100%	93. 3%

表 3　C 均值聚类结果混乱矩阵

Tab. 3　Confused matrix of G-means clustering method

	层云	积雨云	卷云
层云	36	0	0
积雨云	10	30	1
卷云	0	12	19

表 4　FCM 聚类结果混乱矩阵

Tab. 4　Confused matrix of FCM method

	层云	积雨云	卷云
层云	36	0	0
积雨云	10	30	1
卷云	0	6	25

表 5　半监督 FCM 聚类结果混乱矩阵

Tab. 5　Confused matrix of semi-FCM method

	层云	积雨云	卷云
层云	34	2	0
积雨云	1	35	5
卷云	0	0	31

4　小 结

本文使用一种新的描述云内部结构和差异性的特征: DI 特征, 结合纹理特征构建了二维云分类特征空间。基于该二维特征空间, 采用半监督的 FCM 聚类方法对云的自动分类问题进行了积极的探讨和实验。研究结果表明, 基于二维云特征空间的半监督 FCM 聚类方法具有较好的分类准确率, 可有效地完成实际卫星云图的分类。该方法利用隶属度这一概念有效改善了机械区分云类的不足, 引入半监督概念克服了无监督聚类的盲目性。本文也存在一定的不足, 云类样本的选区应更具有代表性和可靠性, 分类特征的选取仍显粗糙, 这些将是后续研究中必须加以改进和提高之处。

参 考 文 献:

[1] Koffler R, Decotiis A G, Rao P K. A Procedure for Estimating Cloud Amount and Height from Satellite Infrared Radiation Data[J]. Mon. Wea. Rev., 1973, 101: 240- 243.

[2] Desbois M, Seze G, Szejwach G. Automatic Classification of Clouds on METEOSAT Imagery Application to High-level Clouds[J]. J. Appl. Meteor., 1982, 21: 401- 402.

[3] Welch R M, Navar M S, Sengupta S K. The Effect of Resolution upon Texture-based Cloud Field Classification[J]. J. Geophys. Res., 1989, 94: 14767- 14781.

[4] 郁凡. 多光谱 GMS 卫星图像气象特征量的提取及其在中尺度数值预报模式中的应用[D]. 南京: 南京大学, 1998.

[5] Haralick R M, Shanmugam K, Dinstein I. Textural Features for Image Classification[J]. IEEE Trans. Syst., Man. & Cybern., 1973(3): 610- 621.

[6] Magurran A E. Ecological Diversity and Its Measurement[M]. New Jersey: Princeton University Press, 1988.

[7] Bezdek J C. A Convergence Theorem for the Fuzzy ISODATA Clustering Algorithms[J]. IEEE Tran. Patter Anal. and Machine Intell., 1980(2): 1- 8.

[8] 梁夷龙, 王松, 夏绍玮, 等. 基于超椭球模糊聚类的人脑磁共振图像分割[J]. 软件学报, 1998(9): 683- 689.

[9] Pal N T. On Cluster Validity for the Fuzzy G-means Model[J]. IEEE Trans. Fuzzy Systems, 1995, 3(3): 370- 379.

[10] Amini M, Gallinari P. Semi-supervised Learning with Explicit Misclassification Modeling[C]//Proceedings of the 18th International Joint Conference on Artificial Intelligence, 2003: 555- 561.

[11] Abdelhamid B, Witold P. Enhancement of Fuzzy Clustering by Mechanisms of Partial Supervision[J]. Fuzzy Sets and Systems, 2006, 157(13): 1773- 1759.

[12] Abdelhamid B, Witold P. Data Clustering with Partial Supervision[J]. Data Mining and Knowledge Discovering, 2006, 12(1): 47- 78.