云分类中逐个修改聚类和模糊聚类分类性能的对比研究

朱亚平[1,2] 刘健文[2] 白 洁[2]

(1. 解放军理工大学气象学院,南京 211101; 2. 航空气象研究所)

提 要:利用卫星图像对各种云型进行识别在大气科学领域具有重要意义,为了深入了解云分类过程中逐个修改聚类和模糊聚类对各种云型的识别能力,采用极轨卫星EOS/MODIS图像资料和静止卫星GMS-5图像资料,在样本采集和特征提取的基础上,选择不同的光谱或纹理特征对两种分类器的分类性能进行测试和对比分析。结果发现,不管采用哪种图像资料,提取哪些特征量,逐个修改聚类的平均分类准确率总体上略高于模糊聚类。但就两种分类器对各种云型的识别能力而言,模糊聚类对低云和高云(如层云、薄卷云、密卷云、卷层云、积云)的分类准确率明显好于逐个修改聚类,而逐个修改聚类对积雨云的分类准确率稍高于模糊聚类。从各类别间混判的情形来看,积雨云和高中低混合云、低云之间及卷云子类之间混判的情形较多,模糊聚类与逐个修改聚类相比,混判的类别增多,相对比例减少。

关键词:云分类 模糊聚类 逐个修改聚类

A Comparative Study on Stepwise Cluster and Fuzzy Cluster in Cloud Classification Techniques

Zhu Yaping[1,2] Liu Jianwen[2] Bai Jie[2]

(1. Institute of Meteorology, PLA University of Science and Technology, Nanjing 211101;
2. Beijing Institute of Aviation Meteorology)

Abstract: In order to profoundly understand abilities of two classifiers—stepwise cluster and fuzzy cluster in the cloud classification techniques, both EOS/MODIS and GMS-5 data set are used, spectral or textural features are drawn from samples randomly to identify various cloud/surface. The results show that the stepwise cluster gives higher accuracies than fuzzy classifier on the whole. With regards to discriminating diverse cloud / surfaces, fuzzy cluster demonstrates its higher accuracies thanstepwiseclusterontheclasseshavingsimilarcharacteristicssuchasstratus, cumulostratusand

收稿日期:2006年4月26日; 修定稿日期:2007年1月4日

cumulus; while stepwise cluster has better capabilities of distinguishing cumulonimbus and surfaces. As far as misclassification of cloud/surfaces, fuzzy cluster tends to show lower accuracies in more misclassified classes.

Key Words: cloud classification fuzzy cluster stepwise cluster

引 言

云作为重要的气象要素,对大气科学的各个领域产生了巨大影响。已有研究表明[1-2],大气环流模式对云参数化非常敏感,更准确的云分型与云辐射作用的参数化同样重要,对云进行参数化时还应充分考虑云型中更细致的子类(例如卷云,密卷云,卷层云)对模式的影响。另外,云与天气气候也存在密切关系,对云进行正确的分类有助于我们对天气气候变化的理解和正确预报。因此,利用卫星图像资料进行云分类就成为一个非常重要的研究课题。

关于云分类的研究,国内外学者已探索了几十年。早期用于分析卫星云资料的自动方法是阈值法[3-4]。阈值法原理简单,计算方便,但阈值缺乏定量统一的描述。因此,大量学者采用图像处理和统计分析相结合的方法对云图进行自动识别。传统的统计方法主要有判别分析和聚类分析[5-6]。1990年代人工智能法的引入对客观云分析产生了深远的影响。有关神经网络[7-8]、最大似然估计[9]和模糊聚类[10]的人工智能方法不断涌现。

从某种意义上讲,各种云分类技术(如判别分析、聚类分析、神经网络等)都是统计模式识别技术,它们最根本的差别体现在分类器的不同。神经网络云分类的处理和分析过程比较庞杂,同时因其系统本身的一些局限性(容易陷入局部极小点;收敛速度慢)使得该方法在业务环境中存在很多困难。而聚类分析运算稳定、操作简便,在业务上更加可行。在国内以往的研究中,模糊逻辑方法仅限于台风云系的分析[11-12],用于其他云型的客观分类还较少。另外,逐个修改聚类大多是对高云、中云、低云等进行大致的分类,缺乏更为细致的子类分析;混合的多层云和卷云仍是云型客观分析中的难点。因此,本文采用不同的光谱或纹理特征向量对逐个修改聚类和模糊聚类的分类器进行对比分析,以深入了解两种分类器在识别不同云型方面的分类性能。

1 分类器

合理的分类器是正确分类的关键,也是云分类的难点。根据分类器的不同,统计模式识别的云分类技术通常可分为监督分类和非监督分类[13]。监督分类技术需要足够的先验知识,要得到合理的分类结果,必须建立所有可能出现的各种云型,样本的质量和数量成为很重要的影响因子。而非监督分类并不需要充当分类依据的"历史资料"作为分类的指导,只根据事物本身的性质来进行分类。因此,本文采用了非监督分类技术,将图像处理和统计方法相结合,建立了逐个修改聚类和模糊聚类的分类器,并根据卫星图像的光谱或纹理特征对两种分类器的分类性能进行分析。将卫星图像资料中获得的 m 个待分类样本组成特征向量$\{x_1, x_2, \Lambda, x_m\}$,每个样本由 n 个光谱或纹理特征描述 $x_i=\{x_{i1}, x_{i2}, \Lambda, x_{in}\}$,$i=1,2,\cdots\cdots,m$。

1.1 逐个修改聚类

(1) 首先给定几个初始凝聚点作为聚类中心 x_i,$i=1,\Lambda,K$,计算 K 个凝聚点的重心(均值)G_i,$i=1,\Lambda,K$;

(2) 按欧式距离公式计算初始聚类中心

两两之间的距离 $d_{G_i,G_j}=[\sum_{t=1}^{n}(x_{it}-x_{jt})^2]^{1/2}$。根据凝聚点间距离的临界值 C 判断各类别间性质的差异性,如果 $\min\{d_{G_i,G_j}\}\geqslant C(i,j=1,\cdots,K)$,说明 G_i 与 G_j 代表性质不同的两类,确为 K 个初始凝聚点;反之,若 $\min\{d_{G_i,G_j}\}<C(i,j=1,\cdots,K)$,则说明 G_i 与 G_j 性质相近,将相应的两类凝聚点合并 $K=K-1$,用两者的重心作新凝聚点。对 $K-1$ 个凝聚点重复上述计算,直至所有初始凝聚点间的最小距离均小于 C 为止;

(3) 将其余 $m-K$ 个未作凝聚点的样本逐个进行归类,即计算样本和聚类中心的欧式距离 $d_{x_j,G_i}(i=1,\wedge,K;j=K+1,\wedge,m)$。按照样本和凝聚点间的距离临界值 R 对样本进行归类,如果 $d_{x_iG_j}>R$ 则此 x_j 为新的聚类中心;反之,若 $d_{x_iG_j}\leqslant R$,则 x_j 归入与它最近的凝聚点那一类 $x_j\in G_i$,使得分解的误差最小,同时重算这类重心,并以此重心为新凝聚点;

(4) 重新检验聚类中心间的距离,如果最小距离有小于 C 的用(2)合并,直至所有凝聚点间距离均大于等于 C。将剩余样本重新按(3)的步骤检验归类。

由于一次归并各凝聚点可能不太稳定,还需按照初始分类的步骤,将各样本从头至尾再逐个进入,归并聚类。若聚类中某个样本进入后与原来分类不同,这两类凝聚点都要重算,当多次逐个进入与上一次分类全同时,聚类过程结束。聚类过程中,初始聚类中心的个数可以从 m 个样本中随机选取,聚类中心的选取只对分类过程产生影响,而对分类结果没有太大影响;凝聚点间距离的临界值以及样本和凝聚点间的距离临界值 $R(C\leqslant R)$均可以通过对训练样本的分类过程进行调整,从而获得稳定的经验值。

1.2　模糊聚类

(1) 首先假定一个初始分划矩阵 $\boldsymbol{U}^0$,矩阵元素满足以下条件:$0\leqslant u_{ij}\leqslant 1;,i=1,\wedge,c;j=1,\wedge,n,\sum_{i=1}^{c}u_{ij};1\leqslant j\leqslant n,\sum_{j=1}^{c}u_{ij}>0;i=1,\wedge,c,u_{ij}$ 反映第 j 个样本 X_j 对第 i 类的隶属关系,称为隶属度,也称置信度,其中 n 表示样本数,而 c 则对应了样本所属的类别数;

(2) 根据初始隶属矩阵计算聚类中心 $V_i=\sum_{i=1}^{m}u_{it}x_{tj}/\sum_{i=1}^{m}u_{it},i=1,\wedge,c;j=1,\wedge,n$,按欧式距离公式 $d_{ij}=\|x_j-V_i\|=[\sum_{i=1}^{n}(x_{jt}-v_{it})^2]^{1/2}$ 计算样本 X_j 与聚类中心 V_i 之间的距离;

(3) 根据 $u_{ij}=(\|x_j-V_i\|^2)^{-1}$ 重新计算分划矩阵 $\boldsymbol{U}$,如果 $\max\{|u_{ij}-u_{ij}^0|\}\geqslant\varepsilon,\varepsilon$ 是任意给定的一个很小的整数(ε 可取 10^{-3},10^{-4} 或 10^{-5} 等),回到(2)根据已得的矩阵 $\boldsymbol{U}$ 算出新的聚类中心,并重新计算样本和聚类中心间的距离,不断调整 V_i 和 $\boldsymbol{U}$,直到隶属矩阵满足 $\max\{|u_{ij}^{p+1}-u_{ij}^{p}|\}<\varepsilon$,则聚类过程结束,所得到的 V_i 和 $\boldsymbol{U}$ 即为最终聚类中心和隶属矩阵;

(4) 根据最终的隶属矩阵对样本进行归类,即将 $\boldsymbol{U}$ 中每一列的元素中最大者取为 1,其它元素均取为 0,实际上是将样本划归到从属程度最大的那一类。

从分类过程来看,模糊聚类与逐个修改聚类的迭代算法类似,它也需要给定初始聚类中心,并分配隶属等级,根据样本到聚类中心的距离最小化对聚类中心和隶属等级不断调整,从而完成对样本的分类。两者间的区别在于,模糊聚类对样本进行归类时采用了软划分,比逐个修改聚类的硬划分更加合理[14],逐个修改聚类是将样本绝对地归到某一类中,每个样本必属于一类;而模糊聚类是将每个样本看作“模糊的”,即一个样本并非绝对地属于哪一类,可能同时具有几个类别的特征,它通过样本隶属于某个类别的可能程度进行划分,更适于边界模糊的系统[10]。有关逐个修改聚类和模糊聚类的详细内容可参看文献[14-16]。

2 对比分析

为了更好地了解两种分类器的分类性能，本文利用极轨卫星 EOS/MODIS 图像资料和静止卫星 GMS-5 图像资料作了试验。随机挑选包括 10 种云/表面类型(见表 1，表 2)在内的极轨卫星样本 306 个($m=306$)，其中，由于积雨云和层云(或雾)的样本很少，因此训练集中仅用了 9 个样本，其他的 8 种类型每类 36 个样本。对于静止卫星，则随机抽取了包括 11 种云/表面类型(见表 3，表 4)在内的静止卫星样本 656 个($m=656$)，其中，由于单层高积云(或高层云)的样本很少，因此训练集中仅用了 16 个样本，其它的 10 种类型每类 64 个样本。由于各种云往往同时出现，形成特征明显的多层云系，为了更客观地描述各种云型的特征，引入几种混合云型作为目标集，中低云指中低混合云，高中低云指高中低混合云。另外，由于在卫星云图上很难区分出高积云和高层云，高积云就表示出现的单层中云。

在样本采集的基础上提取图像的光谱或纹理特征并进行特征分析，根据特征分析结果选择有代表性的光谱或纹理特征。极轨卫星 EOS/MODIS 选择 25 个通道的亮温或反照率的最小值、最大值、标准差；16 个红外通道两两之间亮温差的最小值、最大值、均值、标准差，暂不使用有关反照率比值的特征。静止卫星的光谱特征选择除水汽通道标准差以外的有关亮温或反照率的所有特征，暂不使用有关灰度级的特征；一阶概率特征[17-18]选用红外和可见光通道的能量、熵、水汽通道的逆差距以及 4 个通道的惯量；灰度级差矢量特征[16-17]选用红外和可见光通道的能量、熵、逆差距和惯量，暂不考虑有关水汽通道的特征量，有关内容可以参见文献[19]。将上述极轨卫星的光谱特征组成特征向量 X_1($n=91$)；静止卫星的光谱特征组成特征向量 X_2($n=16$)，光谱特征和一阶概率特征组成特征向量 X_3($n=27$)，光谱特征、一阶概率特征、灰度级差矢量特征组成特征向量 X_4($n=75$)。将几个不同的特征向量分别带入逐个修改聚类和模糊聚类的分类器，对两种分类器的分类性能进行测试和对比分析。

本文以 EOS/MODIS 光谱特征 X_1、GMS-5 光谱和一阶概率特征 X_3 进行逐个修改聚类和模糊聚类为例(见表 1、表 2、表 3、表 4)，表中对角线上的元素表示随机样本经识别后划分到各类的样本数占该类样本总数的百分比，即为该类别的分类准确率，其它元素则表示该类样本被判为其它类别占该类样本总数的百分比，即该类别与其它类别间的混判率。

在利用 EOS/MODIS 样本进行对比试验的结果中(见表 1、表 2)，逐个修改聚类的平均分类准确率略高于模糊聚类。对于各种云/表面类型来说，逐个修改聚类对积雨云的分类准确率达到 88.9%，而模糊聚类仅有 80.56%；对于地表、海表，逐个修改聚类均达到了 100%，而模糊聚类分别为 94.44%、91.67%，对高中低混合云，逐个修改聚类达到 75%，而模糊聚类仅有 61.11%。对低云和高云而言，模糊聚类的分类准确率明显高于逐个修改聚类，层云、薄卷云、密卷云、卷层云、积云的分类准确率模糊聚类高出逐个修改聚类近3%，而层积云的分类准确率逐个修改聚类为 77.78%，模糊聚类却达到了 83.33%。

另外，从逐个修改聚类各类别间混判的情形来看，积雨云和高中低混合云、低云之间及卷云子类之间混判的情形较多；在模糊聚类中仍然存在各类别间的混判情形，但混判的类别增多，相对比例减少，尤其是高中低混合云和低云之间，这是由于模糊聚类对性质相似的类别比较敏感，而各种云类的样本总是或多或少夹杂了其它类型的像素，同时各种云本身的光谱或纹理特征存在一定的不稳定性，所以采用模糊聚类得到这样的结果是合理的。

表 1　采用 EOS/MODIS 光谱特征 X_1 逐个修改聚类结果

	积雨云	层云	薄卷云	密卷云	卷层云	积云	层积云	高中低云	海表	地表
积雨云	88.9							11.1		
层云		75					25			
薄卷云			72.22	8.33	16.67	2.78				
密卷云	5.56			72.22				22.22		
卷层云			8.33	11.11	77.78			2.78		
积云						80.56	19.44			
层积云						22.22	77.78			
高中低云	13.89		11.11					75		
海表									100	
地表										100

(平均分类准确率:81.94%)

表 2　用 EOS/MODIS 光谱特征 X_1 模糊聚类结果

	积雨云	层云	薄卷云	密卷云	卷层云	积云	层积云	高中低云	海表	地表
积雨云	80.56							19.44		
层云		77.8					22.2			
薄卷云			75	5.56	13.88	5.56				
密卷云	2.78			75	8.33			16.67		
卷层云			8.33	2.78	86.11			2.78		
积云						83.33	11.11			5.56
层积云						16.67	83.33			
高中低云	27.78		11.11					61.11		
海表		5.56							94.44	
地表			2.78			5.55				91.67

(平均分类准确率:80.84%)

利用 GMS-5 样本进行对比试验的结果(见表 3、表 4)也表明,逐个修改聚类对积雨云、地表、海表的分类准确率明显高于模糊聚类,但对于积云、层云、层积云、高积云(或高层云)、中低混合云、高中低混合云等,模糊聚类的分类准确率都高于逐个修改聚类;而且,这样的结果和 EOS/MODIS 基本上是一致的,但对高中低混合云的分类却有不同,利用 EOS/MODIS 资料采用模糊聚类对高中低混合云的分类准确率很低,与逐个修改聚类差异较大,这可能和极轨卫星资料分类过程中仅采用光谱特征未引入纹理特征有关。

表 3　采用 GMS-5 光谱和一阶概率特征 X_3 逐个修改聚类分类准确率

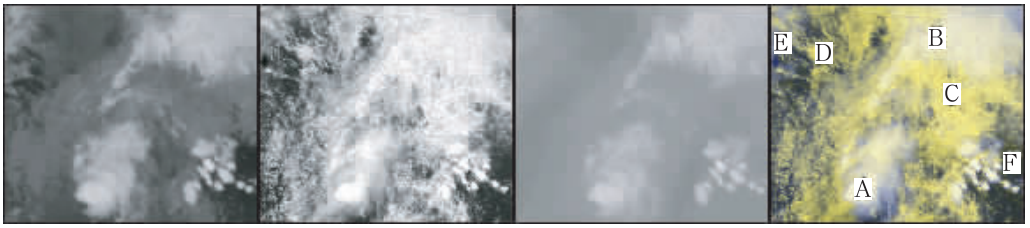
	积雨云	卷云	积云	层云	层积云	高积云	中低云	高中低云	海表	植被	沙地
积雨云	90.63							9.37			
卷云		75	12.5				12.5				
积云			70.31		20.31			9.38			
层云				75	25						
层积云			10.81		71.88	4.69	12.62				
高积云					23.44	54.69	21.87				
中低云			4.69		14.06	9.37	76.56				
高中低云	14.06							85.94			
海表				3.12					96.88		
植被		1.56	6.25							81.25	10.94
沙地		3.13	4.69							7.8	84.38

(平均分类准确率:78.41 %)

表 4 采用 GMS-5 光谱和一阶概率特征 X_3 模糊聚类的分类结果

	积雨云	卷云	积云	层云	层积云	高积云	中低云	高中低云	海表	植被	沙地
积雨云	89.06							10.94			
卷云		67.19	15.63		6.24	1.56	4.69				4.69
积云			71.88		12.5	1.56	12.5			1.56	
层云				79.69	6.25		12.5		1.56		
层积云			1.56	7.82	73.44	7.81	9.37				
高积云			3.13		18.74	64.06	14.07				
中低云			9.36		12.5	4.7	76.56				
高中低云	12.5							87.5			
海表				6.25					93.75		
植被		2.87	6.25							81.25	9.38
沙地		6.25	9.37							10.94	73.44

（平均分类准确率：77.98 %）



(a) (b) (c) (d)

图 1 2002 年 6 月 8 日 03 时 GMS-5 图像
（a）红外一通道图像（b）可见光通道图像（c）水汽通道图像（d）伪彩合成图像

两种分类器的分类效果可以通过以下个例进行说明。图 1(a),(b),(c),(d)分别是 6 月 8 日 03 时位于 29.86°—39.86°N、99.39°—115.39°E 的 GMS-5 红外一通道、可见光通道、水汽通道及伪彩合成图像。借助静止卫星图像动画显示，2002 年 6 月 7 日 03 时在我国西南地区上空为大范围的中低云所覆盖，其中有色调白亮的强对流云团，随着系统发展加强，逗点云系逐渐形成得到发展，6 月 8 日 03 时逗点云系头部的卷云在高空呈明显的辐散形式，尾部的强对流云团也更加明显，这在伪彩合成图像上有清楚的体现，A 处和 B 处对应了明显的高中低混合云，其中有色调更加白亮的积雨云，C 处上空附近覆盖了积云和层积云组成的中低混合云，D 处在红外图像上色调较暗，而可见光图像色调白亮，对应了层积云，E 处附近上空覆盖了薄卷云，F 处则是明显的晴空区。

图 2(见彩页)给出了逐个修改聚类结果和模糊聚类结果。从分类结果看，两者的分类结果大体形式比较一致，但逐个修改聚类可以较好的区分出积雨云，而模糊聚类将积雨云识别为高中低混合云，从图 1 中可以看到，积雨云的色调和高中低混合云还是存在一定差异，因此，逐个修改聚类的结果好于模糊聚类；在层积云和中低混合云的识别方面，逐个修改聚类将层积云识别为中低混合云，但在图 1 中红外图像上可以明显看到 D 处的色调较暗，模糊聚类的结果明显好于逐个聚类。在该例中，高中低混合云、卷云、晴空的识别两者的结果差别不是很大。通过该例可以较好地说明两者的分类性能存在的差异。

3 总 结

本文采用极轨卫星 EOS/MODIS 图像资

料和静止卫星 GMS-5 图像资料,选择不同的光谱或纹理特征向量对逐个修改聚类和模糊聚类的分类器进行了测试和对比分析。结果表明,不管采用哪种图像资料,提取哪些特征量,逐个修改聚类的平均分类准确率总体上略高于模糊聚类。但就两种分类器对各种云型的识别能力而言,模糊聚类对低云和高云(如层云、薄卷云、密卷云、卷层云、积云)的分类准确率明显好于逐个修改聚类,而逐个修改聚类对积雨云的分类准确率稍高于模糊聚类。从各类别间混判的情形来看,积雨云和高中低混合云、低云之间及卷云子类之间混判的情形较多,模糊聚类与逐个修改聚类相比,混判的类别增多,相对比例减少。总的来说,逐个修改聚类适用于类别间差别比较明显的情况,而模糊聚类对类别间相似的情况效果更好。

参考文献

[1] Geleyn, J. A., A. Hense, H. J. Preuss: A comparison of model generated radiation fields with satellite measurements[J]. Contrib. Atmos. Phys., 1982.55: 253-286.

[2] Welch, R. M., B. A. Wielicki Stratocumulus cloud field reflected fluxes: The effect of cloud shape[J]. J. Atmos. Sci., 1984,41: 3085-3103.

[3] Koffler, R., et al. A procedure for estimating cloud amount and height from satellite infrared radiation data [J]. Mon. Wea. Rev., 1973,101: 240-243.

[4] Shenk, W. E., R. T. Holub, and R. A. Neff. A multispectral cloud type identification method developed for tropical ocean area with Nimbus-3 MRIR measurements [J]. Mon. Wea. Rev., 1976,104: 284-291,

[5] Parikh, J. A comparative study of cloud classification techniques[J]. Remo. Sens. Environ., 1977,6: 67-81.

[6] Welch, R. M., K. S. Kuo, B. A. Wielicki Marine stratocumulus cloud fields off the coast of southern California observed using Landsat imagery. Part I: Structural characteristics[J]. J. Appl. Meteor., 1985,27: 341-362.

[7] Key, J. Cloud cover analysis with arctic advanced very high resolution radiometer data 2. classification with spectral and textural measures[J]. J. Geophys. Res., 1990, 95:7661-7675.

[8] Miller, S. W., W. J. Emery. An automated neural network cloud classifier for use over land and ocean surfaces[J]. J. Appl. Meteoro., 1997,36: 1346-1362.

[9] Ebert, E. Analysis of polar clouds form satellite imagery using pattern recognition and a statistical cloud analysis scheme[J]. J. Appl. Meteorol., 1989,28: 382-399.

[10] Baum et al. Automatic cloud classification of global AVHRR data using a fuzzy logic approach[J]. J. Appl. Meteor., 1997,36:1519-1535.

[11] 李俊,周凤仙.气象卫星台风云图的自动识别方法及其应用[J].应用气象学报,1992,3:402-409.

[12] 于波,等.模糊神经网络在台风云系图像识别中的应用[J].气象,1998,22(1):

[13] 蔡元龙.模式识别[M].西安电子科技大学出版社,1986:1-4.

[14] 楼世博,孙章,陈化成.模糊数学[M].北京:科学出版社,1983:120-124.

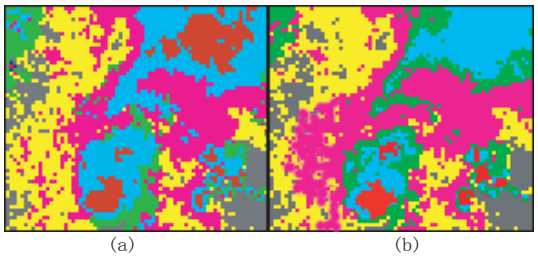
[15] 屠其璞,等.气象应用概率统计学[M].北京:气象出版社,1984:341-343.

[16] 朱亚平,刘健文,白洁.基于 EOS/MODIS 图像资料的多光谱云分类技术[J].海洋科学进展,2004, 22(增刊):109-114.

[17] Haralick, R. M., K. Shanmugam, I. Dinstein. Textural features for image classification. IEEE Trans [J]. Syst. Man Cybern., 1973, 3: 610-621.

[18] Welch, R. M., S. K. Sengupta, D. W. Chen: Cloud field classification based upon high spatial resolution textural features, Part I, Gray level cooccurrence matrix approach[J]. J. Geophys. Res., 1988,93:12663-12681.

[19] 朱亚平,刘健文,白洁.云的光谱和纹理特征统计分析[J].遥感技术与应用,2006,1:18-24.



(a) (b)

图2　2002年6月8日03时GMS-5分类结果

(a) 逐个修改聚类结果 (b) 模糊聚类结果

红色：积雨云；洋蓝色：高中低混合云；绿色：卷云；洋红色：中低混合云；黄色：层积云；灰色：地表