训练样本数目选择对面向对象影像分类方法精度的影响

薄树奎[1)] 丁 琳[2)]

[1)](郑州航空工业管理学院计算机系, 郑州 450015) [2)](中国科学院遥感应用研究所, 北京 100101)

摘 要 面向对象遥感影像分类中的样本选择与基于像素的方法有很大不同,基于统计学理论,研究了面向对象方法的样本数量选择问题。首先,针对面向对象方法的特点,对影像特征空间进行分析,结果表明面向对象方法中要求训练样本的数量可以显著地减少。然后,在遥感影像分类实验中,借助样本数量与波段数目的关系,验证了理论分析的结果。

关键词 分类 面向对象 训练样本 遥感影像

中图法分类号: TP391 **文献标志码**: A **文章编号**: 1006-8961(2010)07-1106-06

The Effect of the Size of Training Sample on Classification Accuracy in Object-oriented Image Analysis

BO Shukui[1)], DING Lin[2)]

[1)](*Department of Computer Science and Application, Zhengzhou Institute of Aeronautical Industry Management, Zhengzhou* 450015)
[2)](*Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing* 100101)

Abstract As opposed to per-pixel classification, the selection of training samples is different in object-oriented method. Based on statistical theory, the number of training samples required in object-oriented classification is studied in this paper. First, feature space analysis of images is implemented in object-oriented classification, which shows that the number of training samples needed for object-oriented classification is much less than that in per-pixel classification. Then, an experiment of remote sensing image classification is carried out to verify the authenticity based on the relations between samples and bands.

Keywords classification, object-oriented, training samples, remote sensing image

0 引 言

面向对象分类方法中的对象是与像素相对应的影像分析实体,是光谱信息类似的相邻像素集合体,其大小由影像分割尺度与影像空间结构决定[1]。面向对象分类方法受到越来越多的研究和应用[2-4],明冬萍[2]基于面向对象方法研究了高分辨率遥感影像信息提取与目标识别。Geneletti[4]则将TM 数据和航空影像结合起来,利用面向对象方法实现土地覆盖分类。面向对象的影像分析技术是一种区别于以往方法的新思路,是一种有效遥感影像分析的方法。

面向对象遥感影像分类方法也是一种监督分类,即需要在已知类别的训练场地上提取各类训练样本,通过选择特征变量、确定判别函数或判别规则,从而把影像中的各个影像对象划归到各个给定的类别。在监督分类中,训练样本的选择是一个关

基金项目:国家自然科学基金项目(40771140);河南省科技攻关计划项目(092102210307);河南省高等学校青年骨干教师资助计划;河南省科技厅基础与前沿技术研究计划(092300410043)

收稿日期:2009-02-25;**改回日期**:2009-09-25

第一作者简介:薄树奎(1976—),男,讲师。2007 年于中科院遥感所获地图学与地理信息系统专业博士学位。主要研究方向为信息提取。E-mail: bsk586@163.com

键步骤,一般情况下,分类所要求的训练样本数量是特征变量维数的函数,随着维数的增加而增加。虽然遥感影像数据携带的信息增多了,但是基于像素的传统方法要求大量训练样本,而面向对象分析方法中的基本处理单元和数据总体都发生了改变,在分类中如何有效地选择训练样本数量以及怎样进行有效的参数估计,是面向对象方法中尚未得到具体讨论的问题。

针对面向对象遥感影像分类方法的特点,分析了影像经过分割后形成影像对象的数据空间特点,对分类中的训练样本数量的选择进行了研究,讨论了训练样本数量对分类精度的影响,指出了面向对象分类中的样本选择数量比通常的选取规则可以大大减少,而且与所分析数据空间的复杂程度相关。

1　遥感分类中的样本数量问题

在遥感影像中,地面任意模式的光谱反射特性可以用一个 n 维的向量代替,该向量的每一维代表着该类模式对相应波段光谱的反应。对于每一种模式,由于测量误差和物理环境变化等原因,同一类模式的特征向量在特征空间绝对不可能聚焦在一个点上,而总是按照一定的概率密度散布在某一空间范围内。根据经验,一般可以认为每一种模式在特征空间中都服从正态分布,并用一个向量和一个方差矩阵来描述,其中向量描述了该模式在特征空间中的位置,而方差矩阵描述了该模式在特征空间中的形状。作为类别统计信息的向量均值和方差,它们通常都是根据训练样本来进行估算。而参数估算的精度与训练样本的数量和特征空间的维数息息相关。在基于像素的方法中,对多光谱影像,由于其维数较少,训练样本的数量相对于特征空间的维数有着较大的比率,因而可以得到较为准确的参数估计值;对于高光谱影像,由于维数的大幅度增加,导致用于参数估计所需的训练样本数量也急剧增加。如果训练样本的数量满足不了特征空间维数增加的要求,则估计出的参数精度就难以保证,比如某些重要的地面覆盖信息,由于所占面积较小,不能提供足够数量的训练样本点,往往不能得到满意的分类结果。在这种情况下,虽然光谱波段数目的增加隐含了更多的分类信息,但由于参数的估计值不够精确,导致分类的结果与理想情况相差很大。最终导致的结果就是,当训练样本数量一定时,随着高光谱影像维数的增加,分类精度“先升后降”,即产生所谓的 Hughes 现象[5-8]:采用最大似然分类时,在样本数一定的情况下分类精度随波段数增加上升到一定程度后开始下降。由于平均分类错误率与许多因素如特定样本的选择、分类器的选择以及假定的概率结构有关,因此分析一般情况下分类错误率与特征数之间的关系比较困难。Hughes 设计了一个实验方法并得出关于这个问题的平均分类错误,清楚地表明在样本数一定的情况下错误率首先随着特征数增加而下降,然后上升。Hughes 的实验说明训练样本数一定的情况下分类精度在某个特征数处达到极大;而对特定的特征维数,训练样本数量不能少于一定值才能保证分类精度。

为了避免特征维数的增加对分类性能的影响,在遥感影像分类中,假设某个地物类别服从正态分布,要选择训练样本对该类别进行表示,结合统计学理论,需要的样本数量可以由下式计算[9]:

$$n = \frac{\sigma^2 z^2}{h^2 + \frac{\sigma^2 z^2}{N}} \tag{1}$$

式中,h 表示指定的置信区间的半宽,σ 是类别标准差,z 是指定的置信水平,N 是类别的大小。对于比较大的类别,如遥感影像中以像素数目表示的较大的类别,式(1)可以近似表示为

$$n = \frac{\sigma^2 z^2}{h^2}$$

在遥感分类应用中,一般采用试探性的方法确定选择训练样本数量,选取规则是每个类别需要的样本数量为数据波段数的 10 ~ 30 倍[10-11],或者在此基础上越多越好。然而,这种规则并不是严格限制的,对于一些简单的判别问题,在基本不影响分类精度的前提下,所需要的样本数量可以大大地减少。在面向对象影像分类方法中,分类处理的对象不是像素而是影像对象,所选择的训练样本也是影像对象,也就是一组像素组成的一个同质区域作为一个单一样本,这样的样本具有单一的光谱性质,因此与面向像素的方法相比,待分析的数据量大大减小。在实际应用中,对于多光谱影像,不需要按照通常的规则选择 10 ~ 30 倍于波段数目的训练样本,更多的样本并不能相应地提高分类精度,下面对这一问题进行分析。

2　面向对象影像分类中的样本数量选择

由前面的计算公式可知,训练样本需要的数量

与置信区间的宽度、置信水平、类别的标准差和类别大小都是相关的,其中与数据本身相关的最重要的就是类内标准差,这里分析类内标准差在面向对象影像分类中的变化情况。面向对象的遥感影像分析方法,首先对原影像进行分割,根据特定的规则将相互邻近的像素合并成影像对象,然后以这些影像对象为基本处理单元进行分析和处理,每个影像对象的光谱值是其中所有像素的平均值。这样,与同一个遥感影像对应的数据在分割前后是不同的,数据集在影像分割前是所有像素值的总体,而在分割后是一定的像素组合后均值的总体,因此,形成了影像分割前后两个不同的数据集,而且这两个数据集的大小和标准差也不相同。

面向对象影像分类中,分割后每个影像对象内的像素均值作为数据点,而且分割过程使用的参数可以控制分割后影像对象的平均大小,类似在一定的分布总体抽取一定数量的样本,然后计算样本均值的分布。假设一定的数据集(一定分布的总体),已知该分布的期望和方差,从这个总体中抽出一部分(m 个)数据,构成一个样本,计算出一个样本平均值,这样有放回的无数次抽选样本,将会产生无数个样本平均数,而且这些样本平均数具有自己的分布形式。在抽样比非常小的情况下,无放回抽样与有放回抽样的误差基本是相同的,可以利用有放回抽样的误差计算公式来代替无放回的情况,所以影像分割后的数据集分布形式可以类似地计算出来。

对任意分布的总体 X,期望为 EX,方差为 DX,有放回抽选样本,容量为 m,设样本均值为随机变量 y,则

$$y = (x_1 + x_2 + \cdots + x_m)/m$$

其中,x_1, x_2, …,x_m 为总体的 m 个有放回抽样,那么 y 的期望为

$$\begin{aligned} Ey &= E((x_1 + x_2 + \cdots + x_m)/m) = \\ &\frac{1}{m}E(x_1 + x_2 + \cdots + x_m) = \\ &\frac{1}{m}(Ex_1 + Ex_2 + \cdots + Ex_m) = \\ &\frac{1}{m} \times m \times (Ex_i) = EX \end{aligned}$$

在上式中,由于是有放回抽样,各抽样的期望 Ex_i 相等,且抽样分布的数学期望等于总体分布的数学期望 EX。

由于 $Dy = Ey^2 - (Ey)^2$,求 Ey^2 得

$$\begin{aligned} Ey^2 &= E((x_1 + x_2 + \cdots + x_m)/m)^2 = \\ &\frac{1}{m^2}E(x_1 + x_2 + \cdots + x_m)^2 = \\ &\frac{1}{m^2}E(x_1^2 + x_2^2 + \cdots + x_m^2) + \\ &\frac{2}{m^2}E(x_1x_2 + x_1x_3 + \cdots + x_1x_m + \\ &x_2x_3 + \cdots + x_2x_m + \cdots + x_{m-1}x_m) = \\ &\frac{1}{m}EX^2 + \frac{2}{m^2} \times \frac{m(m-1)}{2}(EX)^2 = \\ &\frac{1}{m}EX^2 + \frac{m-1}{m}(EX)^2 \end{aligned}$$

所以

$$\begin{aligned} Dy &= Ey^2 - (Ey)^2 = \\ &\frac{1}{m}EX^2 + \frac{m-1}{m}(EX)^2 - (EX)^2 = \\ &\frac{1}{m}DX \end{aligned}$$

由此可得,抽取的样本均值的期望与总体期望相等,方差为总体方差的 $1/m$。而且统计理论表明,不论总体的分布如何,只要样本容量 m 足够大(大于 30),样本均值的分布总会趋向于正态分布。在面向对象影像分类中,一般都能保证实际分割的影像对象足够大,可以包含从几十到上百个像素。因此,样本均值数据的分布形式可以很好地近似正态分布,这样,训练样本数量计算公式的前提条件可以得到满足。如果分割后的影像对象平均大小为 m 个像素,即从总体中抽取样本容量为 m 的样本并求均值,根据式(1),数据集的类内方差与训练样本数量成正比例关系,对所有均值组成的数据集进行分类,由于其类内方差为总体方差的 $1/m$,所要求训练样本的数量大大减小了。因此,在面向对象分类中,得到同样的分类精度,所需要的训练样本数量也可以明显减少,下面以遥感影像分类实例对所讨论的结果进行验证。

3 实验结果

前面从理论上讨论了针对面向对象影像分类方法的样本数量选择问题,在遥感影像分类中,训练样本的数量通常不是利用公式确定,而是采用人们长期总结出来的简单方法,即选择 10 ~ 30 倍于影像波段数目的训练样本。在面向对象遥感影

像分类中,虽然影像波段数目不变,但与基于像素方法相比数据集的复杂程度减小了,因此要求训练样本的数量也减少了,下面以实验对这一结果进行验证,并进一步讨论训练样本数量和波段数目的关系。

实验数据是一幅 TM 遥感影像,如图 1 所示。该影像是德国 Definiens Imaging 推出的 eCognition 软件附带的实例数据[1],其中有林地、草地、不透水表面和水体 4 个地物类别,“不透水表面”实际就是林地、草地和水体之外的其他类别。而且实例中给出了各个类别的地表真值训练样区,本实验将这些训练样区作为分类结果的检验区域,并结合人工解译进行精度评价。TM 遥感影像有7 个波段,将这 7 个波段进行不同波段数目的组合,波段组合分别为:4,24,234,2345,23456,123456,1234567。采用面向对象方法分别对这些组合后的数据进行分类,其中影像分割后得到的对象平均大小约为 100 个像素,因此满足式(1)及第 3 节讨论的 m 值的理论要求。对于每个影像(即不同数目的波段组合的数据)分别选择不同数量的样本,然后分析样本数量和波段数目对于分类精度的影响。面向对象方法中的样本选择也是基于影像对象的,与基于像素的方法类似,在原影像上基于像素选择训练样本,与此像素对应位置上的影像对象则作为面向对象方法中的训练数据。图 2—图 5 是在不同波段数的影像中各个类别分类精度随样本数量增长而变化的曲线图,各图中(a)是基于影像对象的方法,作为比较,(b)给出了基于像素方法的曲线图。

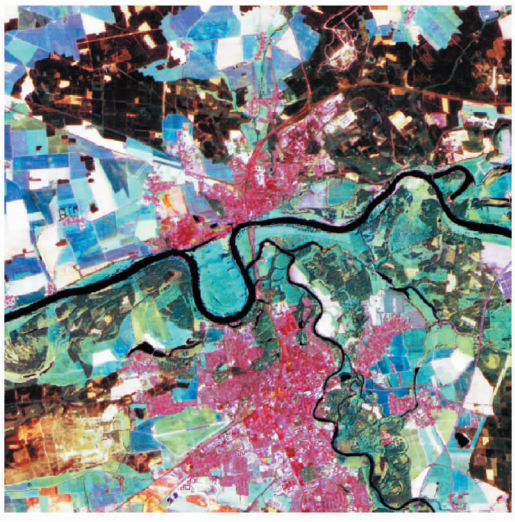

图 1　TM 影像

Fig. 1　TM image

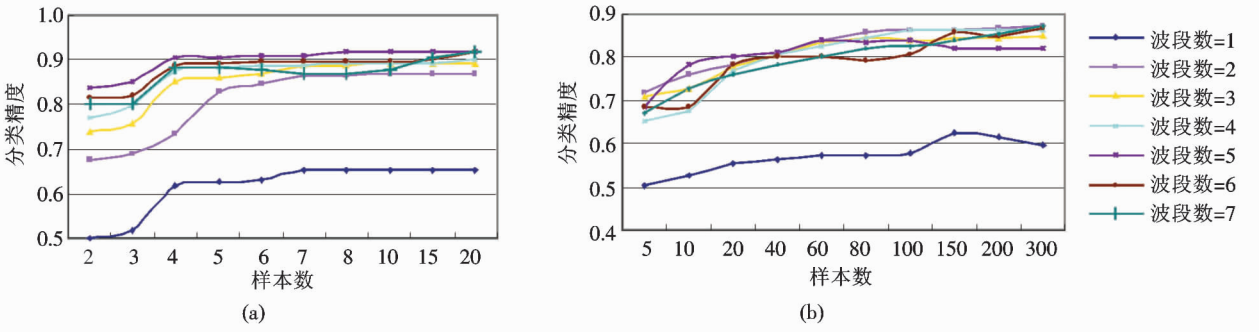

图 2　林地类别的分类精度随样本数的变化

Fig.2　Variation curves of woodland class accuracy with the number of samples

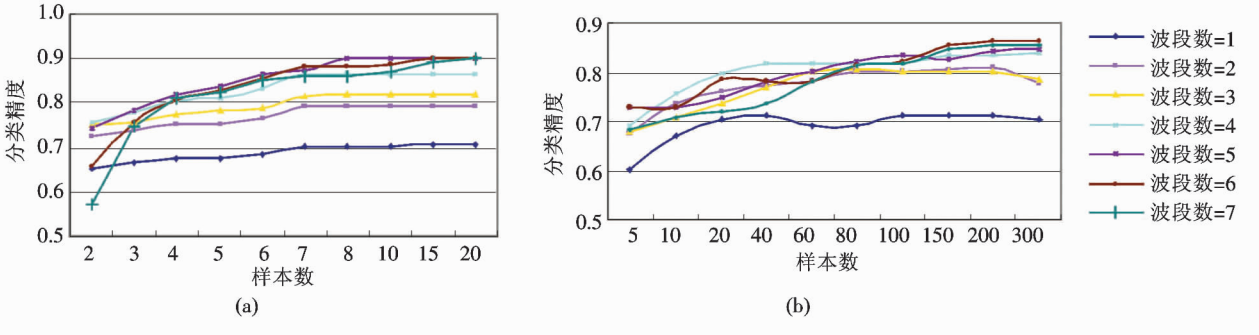

图 3　草地类别的分类精度随样本数的变化

Fig.3　Variation curves of grassland class accuracy with the number of samples

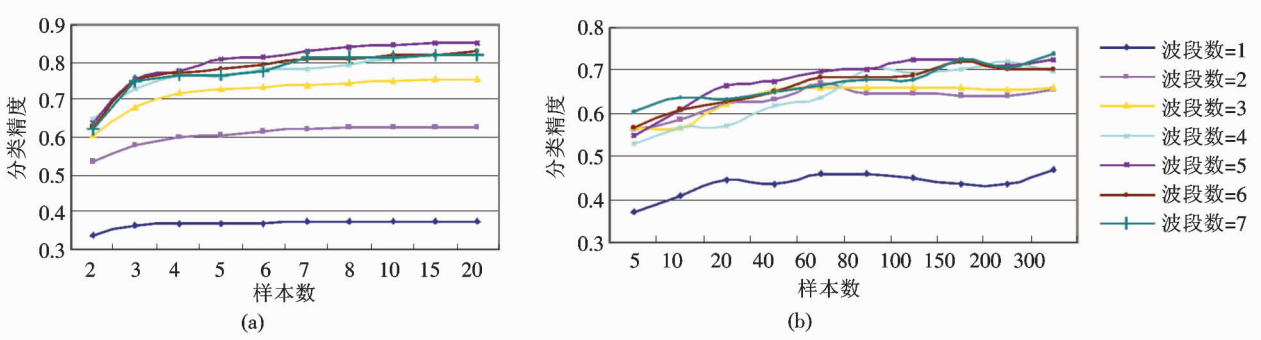

图 4 不透水表面类别的分类精度随样本数的变化

Fig.4 Variation curves of impervious class accuracy with the number of samples

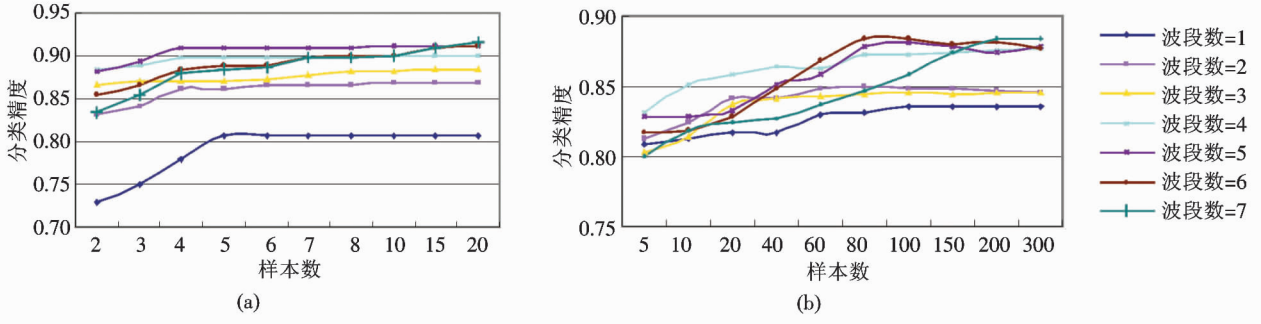

图 5 水体类别的分类精度随样本数的变化

Fig.5 Variation curves of waterbody class accuracy with the number of samples

从图 2—图 5 中可以看出,各类别的分类精度随样本数的变化趋势是近似的,总体上的趋势是随着样本数量的增加,各个类别的分类精度在开始时期上升较快,随着样本数量的增加逐渐放缓,最后都进入一个平稳状态,分类精度在样本数量继续增加时不再有明显提高。由于各个类别数据分布形式的复杂程度不同,进入平稳阶段的起点也不同。例如,林地类别在训练样本数目为 4 ~5 时,分类精度都开始进入平稳期;草地和不透水表面进入平稳期比较慢,大都在训练样本数目为 8 时曲线开始变得平直,由于这两个类别的类内方差较大,根据式(1),要达到同样的精度需要更多的训练样本;而水体类别类内方差小,分类精度随着样本数量的增加,很快进入平稳阶段,而且水体在较少的波段数目时(2 个)分类精度明显高于其他类别,说明了数据的复杂程度对样本数目选择存在显著的影响。各类别的分类精度随样本数的变化情况图表明,在面向对象遥感影像分类中,与基于像素的方法相同,即分类精度随着训练样本数目的增加而增加,当样本数量达到一定值,分类精度基本保持不变;但是面向对象方法比基于像素的分类方法在训练样本的数量上要求少得多,这也证明了前面所讨论的统计理论结果,在实际分类中,面向对象方法只需要 2 ~3 倍于影像波段数目的训练样本就可以达到高而且稳定的分类精度。在分类结果中,当波段数量大于 5 时,4 类地物分类精度均出现普遍下降,原因是开始样本数量少于 2 ~3 倍的波段数目,使得分类精度下降,但随着样本数量的增加(达到 20),4 类地物分类精度又普遍升高(如图 2—图 5)。

从实验中可以看出,面向对象方法的分类精度随训练样本数量的变化与基于像素的方法相似,但是,在面向对象分类方法中,选择训练样本的数量比基于像素的方法明显地减少,样本数量在 2 ~3 倍于波段数目时分类精度就能达到较高的水平,并进入一个平稳阶段。

4 结 论

从理论上讨论了面向对象方法中的训练样本数量选择问题,给出了面向对象影像分类中选择训练样本数量的理论依据,并通过 TM 影像分类实验,进一步说明了训练样本数对分类精度的影响,指出了

面向对象分类中的样本数量比通常的选取规则可以大大减少,而且与所分析数据空间的复杂程度相关。另外,影像分割的结果与分割参数相关,不同的参数会产生不同的影像对象数据集,因此样本数与波段数的关系在不同的分割尺度会有所不同。实验中采用的是最近邻分类算法,不同的算法对样本的要求有所区别,分类精度也不同,但无论哪种分类算法的分类精度都与数据分布形式有关,因此在面向对象影像分类中,采用什么分类算法都有类似的样本数目选择问题。

参考文献(References)

[1] Baatz M, Benz U, Dehghani S, et al. Ecognition Professional User Guide[EB/OL][2009-02-25]. http://www.definiens imaging.com.

[2] Ming Dongping, Luo Jiancheng, Shen Zhanfeng, et al. Research on information extraction and target recognition from high resolution remote sensing image[J]. Science of Surveying and Mapping, 2005, 30(3): 18-20.[明冬萍,骆剑承,沈占锋,等. 高分辨率遥感影像信息提取与目标识别技术研究[J]. 测绘科学, 2005, 30(3): 18-20.]

[3] Chen Yunhao, Feng Tong, Shi Peijun, et al. Classification of remote sensing image based on object oriented and class rules [J]. Geomatics and Information Science of Wuhan University, 2006, 31(4): 316-320.[陈云浩,冯通,史培军,等. 基于面向对象和规则的遥感影像分类研究[J]. 武汉大学学报(信息科学版), 2006, 31(4): 316-320.]

[4] Geneletti D, Gorte B G H. A method for object-oriented land cover classification combining Landsat TM data and aerial photographs[J]. International Journal of Remote Sensing, 2003, 24(6): 1273-1286.

[5] Hughes G F. On the mean accuracy of statistical pattern recognizers[J]. IEEE Transactions on Information Theory, 1968, 14(1): 55-63.

[6] Foody G M, Arora M K. An evaluation of some factors affecting the accuracy of classification by an artificial neural network[J]. International Journal of Remote Sensing, 1997, 18(4): 799-810.

[7] Foody G M, McCulloch M B, Yates W B. The effect of training set size and composition on artificial neural network classification [J]. International Journal of Remote Sensing, 1995, 16(9): 1707-1723.

[8] Shahshahani B M, Landgrebe D A. The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon[J]. IEEE Transactions on Geoscience and Remote Sensing, 1994, 32(5): 1087-1095.

[9] Foody G M, Mathur A. The use of small training sets containing mixed pixels for accurate hard image classification: Training on mixed spectral responses for classification by a SVM[J]. Remote Sensing of Environment, 2006, 103(2): 179-189.

[10] Piper J. The effect of zero feature correlation assumption on maximum likelihood based classification of chromosomes[J]. Signal Processing, 1987, 12(1): 49-57.

[11] Van Niel T G, McVicar T R, Datt B. On the relationship between training sample size and data dimensionality: Monte Carlo analysis of broadband multi-temporal classification[J]. Remote Sensing of Environment, 2005, 98(4): 468-480.