

文章编号: 1001 - 9014(2008)02 - 0123 - 06

基于支持向量机的高光谱遥感图像分类

谭 琨, 杜培军

(中国矿业大学 地理信息与遥感科学系, 江苏 徐州 221116)

摘要:多数传统分类算法应用于高光谱分类都存在运算速度慢、精度比较低和难以收敛等问题. 本文从支持向量机基本理论出发建立了一个基于支持向量机的高光谱分类器, 并用国产 OM IS 传感器获得的北京中关村地区高光谱遥感数据进行试验, 分析比较了各种 SVM 核函数进行高光谱分类的精度, 以及网格搜寻的方法来确定 C 和 γ 的值, 结果表明 SVM 进行高光谱分类时候径向基核函数的分类精度最高, 是分类的首选. 并且与神经网络径向基分类算法以及常用的最小距离分类算法进行比较, 分类的精度远远高于 SVM 分类算法进行分类的结果. SVM 方法在高光谱遥感分类领域能得到广泛的应用.

关键词:高光谱遥感; 支持向量机; 分类
中图分类号: TP751.1 文献标识码: A

HYPER SPECTRAL REMOTE SENSING IMAGE CLASSIFICATION BASED ON SUPPORT VECTOR MACHINE

TAN Kun, DU Pei-Jun

(Department of Remote Sensing and Geographical Information Science, China University of Mining and Technology, Xuzhou 221116, China)

Abstract: Some traditional algorithms used for hyperspectral remote sensing image classification have some problems such as low computing rate, low accuracy and hard for convergence. According to SVM theory, the classification model based on SVM was constructed. By experimenting with hyperspectral image of 64 bands captured by OM IS sensor, the classification accuracy of SVM using different kernel function was analyzed, and the values of C and γ were gained by grid researching. The results indicate that the radial basis kernel function of SVM has the highest accuracy and it can be well used for hyperspectral remote sensing image classification. SVM classifier has more advantages in the classification in contrast with radial basis function neural network classifier and Minimum Distance Classifier (MDC).

Key words: hyperspectral remote sensing; support vector machine (SVM); classification

引言

随着对地观测技术的发展, 各种各样的遥感卫星不断发射成功并源源不断地提供遥感信息. 中国科学院上海技术物理研究所自主设计了高光谱传感器 OM IS、PH 1 等并已经获取了大量航空高光谱影像数据. 高光谱数据的维数高、数据量大、数据不确定性和小样本分类等特点, 是高光谱数据处理的关键和难点, 需要遥感信息科学、模式识别、计算智能等多学科理论和技术的交叉集成.

支持向量机 (Support Vector Machine, SVM) 作

为一种最新的也是最有效的统计学习方法, 近年来成为模式识别与机器学习领域一个新的研究热点. SVM 具有小样本学习、抗噪声性能、学习效率高与推广性好的优点, 能够用于解决空间信息处理分析领域的遥感影像处理. 遥感图象分析与处理是 SVM 应用一个热门的研究方向^[1], 一些主要应用如土地利用分类^[2]、混合象元分解^[3]、遥感影像融合^[4]、多光谱/高光谱遥感分类等^[5]. 目前针对高光谱数据进行分类还是仅仅局限于传统的分类方法, 不但运算速度慢, 分类精度低, 而且出现了严重的 Hughes 现象. 而在高光谱遥感分类中 SVM 具有明显的优越

收稿日期: 2007 - 03 - 12, 修回日期: 2007 - 10 - 10

基金项目: 国家自然科学基金 (40401038)、江苏省自然科学基金创新人才青年学术带头人 (BK2006505) 和国家 863 高技术发展计划 (2007AA12Z162) 资助项目

作者简介: 谭 琨 (1981-), 男, 湖南祁东人, 博士生, 从事高光谱遥感图像处理、支持向量机和神经网络方面的研究.

Received date: 2007 - 03 - 12, revised date: 2007 - 10 - 10

性,因此 SVM 应用被归纳为高光谱遥感分类最重要的进展之一.但总体来看,国内对 SVM 应用于高光谱分类的研究还处在起步阶段.本文研究了支持向量机在高光谱遥感图像分类中的应用,建立了一种基于支持向量机的高光谱遥感图像分类模型.实验结果表明,该分类器与神经网络、最小距离分类器相比具有一定的优势,是一种有效的高光谱遥感图像分类方法.

1 支持向量机分类器

1.1 SVM 的特点和理论基础

支持向量机 (Support Vector Machines, SVM) 是一种基于统计学习理论的机器学习算法,采用结构风险最小化 (Structural Risk Minimization, SRM) 准则,在最小化样本误差的同时缩小模型泛化误差的上界,从而提高模型的泛化能力.不同于一些机器学习算法通常采用经验风险最小化 (Empirical Risk Minimization, ERM) 准则,统计学习理论提出了一种新的策略:将函数集构造为一个函数子集序列,使各个子集按照 VC 维的大小排列;在每个子集中寻找较小经验风险,在子集间折衷考虑经验风险和置信范围,取得实际风险的最小,这种思想称作结构风险最小化^[6].

支持向量机的基本的数学形式是^[7]

$$\min_{w, b} (w, b) = \frac{1}{2} (w \cdot w) \quad (1)$$

约束条件:

$$y_i [(w \cdot x_i) + b] \geq 1 \quad i = 1, \dots, n \quad (2)$$

引入拉格朗日乘子 α_i 上式求解方程为

$$\min_{w, b} L(w, b, \alpha) = \frac{1}{2} (w \cdot w) - \sum_{i=1}^k \alpha_i \{ y_i [(w \cdot x_i) + b] - 1 \} \quad (3)$$

约束条件:

$$y_i ((w \cdot x_i) + b) \geq 1, \quad i = 1, 2, \dots, n \quad (4)$$

对 w, b 求偏导,得到

$$\sum_{i=1}^k \alpha_i y_i = 0, \quad w = \sum_{i=1}^k \alpha_i y_i x_i \quad (5)$$

将式 (5) 代入式 (3), 得到

$$W(\alpha) = \sum_{i=1}^k \alpha_i - \frac{1}{2} \sum_{i=1}^k \sum_{j=1}^k \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \quad (6)$$

求解得到最优解 $w = \sum_{i=1}^k \alpha_i y_i x_i$, 取任一

$\alpha_i \neq 0$, 可求出 b 在结果中, 大部分 α_i 为 0, 将 $\alpha_i \neq 0$ 的样本称为支持向量.

对于高维空间, 如果用内积 $K(x, x)$ 代替最优

分类面中的点积, 就相当于把原特征空间变换到了某一新的特征空间, 此时的优化函数

$$\max W(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j K(x_i \cdot x_j) \quad (7)$$

约束条件:

$$\sum_{i=1}^n y_i \alpha_i = 0 \quad (8)$$

$$\alpha_i \geq 0, \quad i = 1, \dots, n \quad (9)$$

α_i 为与式 (3) 中每个样本对应的 Lagrange 乘子. 这是一个不等式约束下二次函数寻优的问题, 存在唯一解, 而且解中将只有一部分 (通常是少部分) α_i 不为零, 对应的样本就是支持向量.

求解上述问题后得到的最优分类函数是

$$f(x) = \text{sgn}[(w \cdot x) + b] \\ = \text{sgn} \left[\sum_{i=1}^n \alpha_i y_i K(x_i, x) + b \right] \quad (10)$$

式 (10) 中的求和实际上只对支持向量进行 b^* 是分类阈值, 可以用任一支持向量求得, 或通过两类中任意一对支持向量取中值求得. 其中核函数 $K(x_i \cdot x)$ 可以有多种形式:

(1) 线性核

$$K(x_i, x) = (x_i \cdot x) \quad (11)$$

(2) 多项式核

$$K(x_i, x) = (x_i \cdot x + 1)^d \quad (12)$$

d 是自然数.

(3) RBF 核 (Gaussian 径向基核):

$$K(x_i, x) = \exp \left[-\frac{x \cdot x_i}{2} \right], \quad \sigma > 0 \quad (13)$$

(4) Sigmoid 核

$$K(x_i, x) = \tanh(v(x_i \cdot x) + c) \quad (14)$$

S 是 Sigmoid 函数, a, v 是某些常数, 一般根据不同分类体系进行确定.

SVM 的实质在于: 首先通过非线性变换将输入空间变换到一个高维空间, 然后在这个高维空间中进行线性回归, 求取最优线性分类面, 而这种非线性变换是通过定义适当的内积函数实现的.

1.2 SVM 多类分类器

SVM 是一种两类分类器, 如何将 SVM 有效地推广到多类分类是当前的研究热点. 多类 SVM 分类器的构造方法一般包括两种策略: (1) 构造一系列两类 SVM 分类器, 每个分类器用于识别其中两个类别, 并将它们的判别结果以某种方式组合起来实现多类分类; (2) 将多个分类面的参数求解合并到一个最优化问题中, 通过求解该最优化问题

“一次性”地实现多类分类.对于这两类方法,第二类方法对于最优化问题求解过程中的变量远远多于第一类方法,而且进一步导致训练速度慢,分类精度也比较差.因此现有大多数方法均属于第一类,这里采用第一类方法中的“1-v-1”进行多类问题的分类.

针对高光谱数据的高维数、多特征、数据量大等特点,这里选取了“1-v-1”(one against one)方法^[8].如果有 k 类数据,选取第 i 类数据和第 j 类数据构造一个分类器(其中 $i < j$),这样设置 $k(k-1)/2$ 个分类器.对于第 i 类和第 j 类需要解决下述两类分类问题:

$$\min_{a^i, b^i} \frac{1}{2} (w \cdot w) + C \sum_{i=1}^n \xi_i \quad (15)$$

训练点也不再满足约束条件 $y_i((w \cdot x_i) + b) \geq 1 - \xi_i, \xi_i \geq 0$,引入一个松弛变量 ξ_i ,使决策面的约束变为

$$y_i((w \cdot x_i) + b) \geq 1 - \xi_i, i = 1, \dots, n \text{ if } y_i = i \quad (16)$$

$$y_i((w \cdot x_i) + b) \leq -1 + \xi_i, i = 1, \dots, n \text{ if } y_i = j \quad (17)$$

用投票法解决上述问题:若 $\text{sgn}(w \cdot x_i + b^i) = 1$,判断 x 属于第 i 类,第 i 类的票数加 1;反之,第 j 类票数加 1.最后 x 归为拥有最多票数的类.这种投票法称为最大占优(Max Wins)法.但如果两类具有同样的票数,可以选择索引值较小的类.

通过这种方法,来解决 SVM 的多类分类问题是一种比较常用的方法,而且与 1-v-R 以及 DAG_SVM 的多类分类方法相比,当大量数据的时候,投票法能够得到比较高的分类精度^[9].

1.3 SVM 分类器的参数选择

在参数选择上面,采用核函数将实际问题转换到高维空间, C 和 γ 是必备的两个参数,分别为惩罚系数参数和间隔,其取值好坏直接影响到分类精度.在这里采用一种基于交叉验证的网格搜寻的方法来确定 C 和 γ 的取值,用参数选择模型 Grid py 来搜索 C 和 γ 的取值^[10].在分类图像的精度评价上面,根据实际地物对原始图像数字化,把数字化结果视为真实图像分类结果,以此为标准,对 SVM 分类的图像进行精度分析对比.

2 分类方法

2.1 分类数据的预处理

首先对于高光谱数据进行特征提取.一般的遥感图像分类中的光谱维特征提取算法一般有,代数运算法、导数法、变换法.这里选择了均值,标准差,

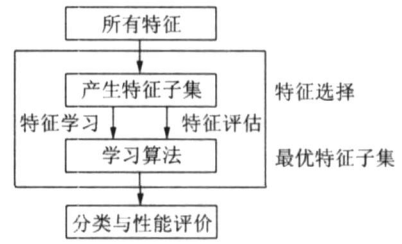


图1 特征提取过程
Fig.1 Features extraction process

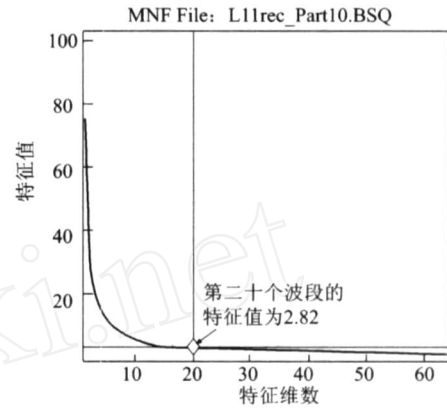


图2 MNF 变换结果
Fig.2 Results of MNF transformation

K-L 变换等特征提取算法.首先选取图片的感兴趣区域,分为 4×4 的窗口进行基本变换.关于纹理窗口大小的选择是经过多次检验的结果,分得太小不但训练速度大大下降,而且结果没有明显变化,分得太大精度太低,所以取了中间结果 4×4 的区域.特征提取流程见图 1.

为了选取信息量大的波段,应用 MNF(Minimum Noise Fraction)变换对图像进行预处理.MNF 变换可以判定图像数据内在的维数(即波段数),分离数据中的噪声,减少随后处理中的计算需求量,其本质上是两次层叠的主成分变换.经过 MNF 变换以后,根据图 2 所示的特征值大小,取得前 20 个特征比较明显的变换后的数据层进行实验.

2.3 基于 SVM 的高光谱分类模型构建

经过多次试验和调整,设置了适合高光谱遥感分类的流程图.主要过程:特征选择、优化与提取;数据归一化处理;选择 C 、 γ 等参数;选择感兴趣区域, SVM 参数的优化, SVM 分类处理,分类后处理等.分类流程见图 3.

3 实验结果与分析

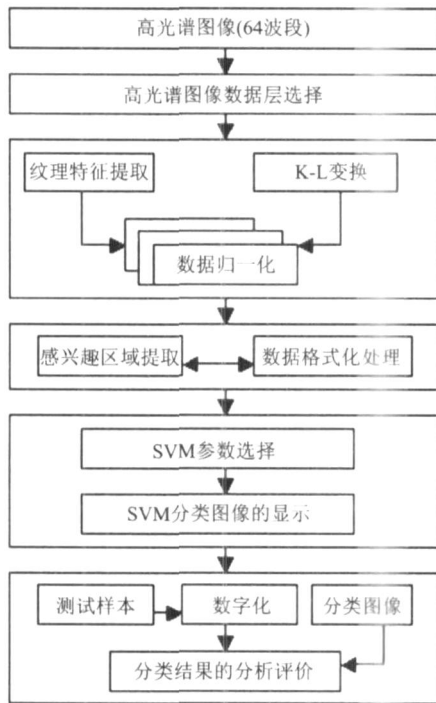


图 3 分类流程图

Fig 3 Experiment process flow chart

3.1 实验数据

实验数据为北京昌平地区的 OM IS 高光谱影像,经过多幅图像的实验,并选其中一幅有代表性的图像(经纬度: 40°10'28.812"N, 116°15'14.7204"E)进行说明,其大小为 512 行, 536 列,共 64 个波段,波段间隔 10nm,飞行高度 1000m。从图 4 可以看出,绿色区域表明该地区为茂盛的植被,黑色区域表示鱼塘的分布区域,黄色的是成熟的稻谷,居民区在影像上面为蓝色,还有一些刚刚种上种子的土地。



图 4 实验区地物分布概况 (36 波段 (0.81μm), 23 波段 (0.68μm), 11 波段 (0.56μm))

Fig 4 Distribution survey of the ground object in testing area (R: band 36 (wavelength 0.81), G band 23 (wavelength 0.68), B band 11 (wavelength 0.56))

表 1 不同核函数的分类结果

Table 1 Classification results of different kernel functions

核函数	SVM 分类精度
高斯径向基核函数 $K(x, x_i) = \exp\left(-\frac{\ x - x_i\ ^2}{2}\right)$ 参数: $\sigma = 0.5$, $c = 8$	76.90%
Sigmoid 核函数 $K(x, x_i) = \tanh[v(x \cdot x_i) + c]$ 参数设置无法达到分类效果	11.65%
线性函数 (Linear) $K(x, x_i) = x \cdot x_i$	66.72%
多项式核函数 $K(x, x_i) = [x \cdot x_i + 1]^d$ 参数: $d = 0.5$	66.22%

表 2 高斯径向基核函数 SVM 分类混淆矩阵

Table 2 Classification confusion matrix of SVM based on RBF kernel function

	C1	C2	C3	C4	C5	列精度
C1	36551	5600	6449	1653	9591	61.08%
C2	5310	23704	1398	1940	1857	69.29%
C3	1410	5725	49852	1474	6515	76.72%
C4	766	1126	50	10131	810	78.64%
C5	4827	2005	1291	770	81339	90.14%
行精度	74.81%	62.12%	84.44%	63.45%	81.25%	76.90%

3.2 训练样本选择

样本选取原则要充分考虑到各种地物的光谱结构和纹理特征,在本研究中,样本分别为 C1:草地, C2:农作物, C3:裸地, C4:水体, C5:建筑物,占实验区数据总量的 2.16%。

3.3 SVM 的不同核函数高光谱遥感图像分类

不同的核函数对应不同的特征映射,从而所学习得到的 SVM 分类超平面也具有不同的特点和能力。核函数选择在 SVM 应用中非常重要,但到目前为止,还没有一个公认比较好的选择标准。为了比较不同核函数的识别效果,这里采用不同的核函数进行实验,实验结果如表 1 所列。

从实验结果来看,选用高斯 RBF 核函数的支持向量机取得了较好的识别结果。与传统的高斯 RBF 神经网络相比,支持向量就充当了聚类中心的角色,当设定的方差较大时,支持向量数量减少;而当方差较小时,支持向量的数目反而增加。所不同的是,支持向量机中的这些聚类中心及其权重都是算法自动寻找到的;而在 RBF 神经网络中,它们需要基于启发式知识进行寻找。在精度评价上面,对原始图像数字化,把数字化结果视为真实图像分类结果,以此为标准,对 SVM 分类的图像进行精度分析对比。

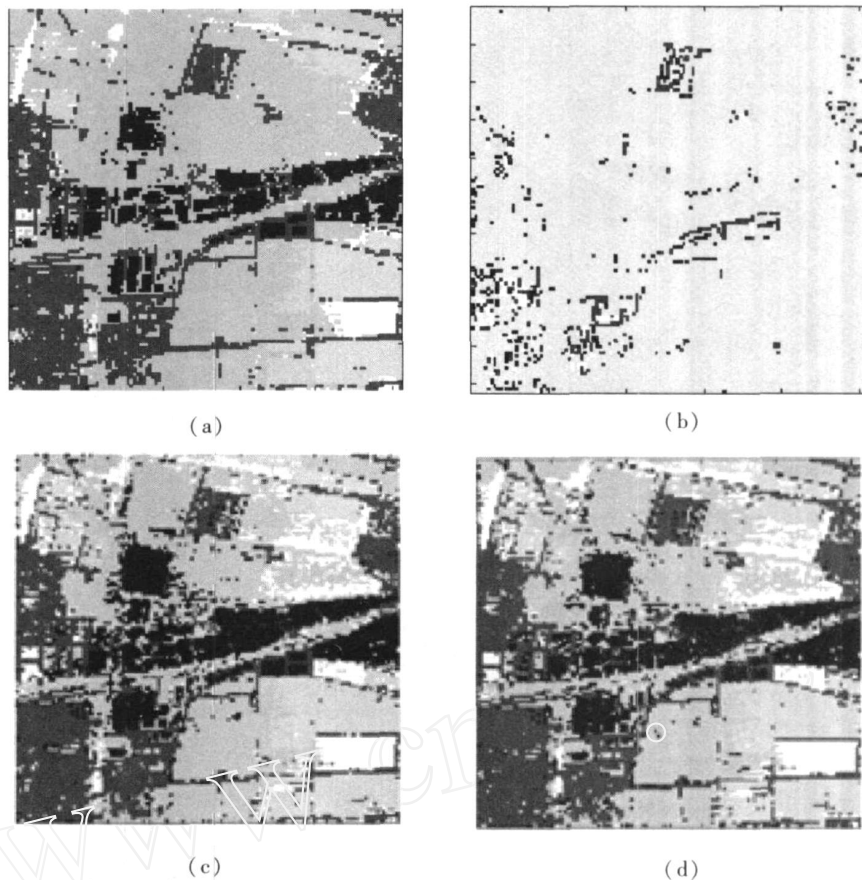


图 5 采用不同核函数的 SVM 分类结果 (a) RBF核函数的分类图 (b) Sigmoid核函数的分类图 (c) 多项式核函数 (d) 多项式和线性核函数分类图

Fig 5 Classification results of SVM classifier using different kernel functions (a) RBF kernel function (b) Sigmoid kernel function (c) polynomial kernel function (d) linearity kernel function

从图 5 (a)中可以看出对于水体的分类结果非常好,在设置合适的参数情况下能得到分类较好的分类精度,由于草地分布的密集和稀疏情况下,所以在草地不密集的区域能够看出土层的情况,建筑物在遥感影像中的光谱和左边角的水体非常相似,出现了水体边缘分成建筑物的情况.成熟稻田和未成熟的也比较难以区分.总体精度还是非常高的.高斯径向基核函数 SVM 分类混淆矩阵见表 2 所列.

对于 Sigmoid 核函数,它与传统的两层神经网络也有着类似的联系,所不同的是,在支持向量机中,隐层单元的数目及其权重都是算法自动寻找的;而在神经网络中,需要通过多次实验或者启发式知识来确定.此外,由于 Sigmoid 函数只对某些 C 和 η 值满足 Mercer 条件,这使它的应用受到了很大的限制.特别对于这里的高光谱数据基本得不出期待的结果.分类结果见图 5 (b).与 RBF 函数分类比起来精度明显下降,草地那块和水体基本混淆在一

起.从分类的精度来看根本不适合分类处理.

而对于多项式核函数来说,由于其具有较明确的物理意义,计算相对比较简单,所以训练用时也较多.线性核函数为多项式函数的特殊情况,分类结果还是不错的.对于水体边缘他们不能很好的进行分类,也很难区分成熟和未成熟稻田.和 RBF 核函数相比也不能分出水体边缘.分类结果见图 5 (c)和图 5 (d).

3.4 神经网络以及最小距离分类结果

在神经网络分类中,一般采用 BP 算法进行分类,但是效果不够理想,特别对于高光谱数据的不确定性,高维数等.为了与 SVM 进行更好的比较,选用了径向基神经网络 (BRFNN) 进行分类.选择同样的训练样本进行分类,分类难点主要是在居民区附近,如图 6 所示,左下角的居民区附近多种地物混杂在一起,比如有水体、草地、建筑物等,这样使得局部地区的精度不够.特别是,当样本足够大的时候,其运算速度非常慢,所以在此应用了小样本学习的思想.



图6 RBFNN 分类结果图

Fig.6 The classification result based on RBFNN

然后又进行了多次分析实验,当 RBFNN 进行高光谱数据分类时候,样本数据需占分类数据的 10% 以下,否则会出现内存异常现象.其精度只有 62.16%.分类结果如图 6 所示.

最小距离法分类是监督分类里面一种常用的方法.其理论基础是把要分类的像素按照最小距离归于一类.但是在这个试验中,它的试验精度和结果低于 RBF 神经网络.其精度只有 40.88%.分类结果如图 7 所示.

4 结语

本文针对传统遥感图像分类方法的缺点,在分析 SVM 分类方法的基础上,对 SVM 函数表达式的核函数进行了分析,而且针对高光谱数据的小样本学习,构建了一种基于支持向量机的高光谱遥感图像分类模型.实验结果表明这种分类器结构简单、容易训练、收敛速度快、具有很高的分类精度,与神经网络方法和最小距离分类方法相比具有一定的优势,是一种有效的遥感图像分类方法.

基于 SVM 的高光谱遥感图像分类在以下方面还有待进一步深入研究:第一,选取合适的纹理特征和其他形状、领域等特征集成起来进行分类,有利于分类精度提高和速度提升.第二,在进行 SVM 分类时候,主要是核函数的选取和参数搜索算法,对于常用的核函数进行了实验,尽管径向基核函数的精度比较高,但是其对硬件要求也随之提高,而且牺牲了大量训练时间为代价.参数选取的时候这里采用的基于交叉验证的网格搜寻的方法,进一步提高参数选取的速率以及参数准确性值得深入研究.第三,多类分类上面在投票算法精度提高以及对于 SVM 核函数的选择上面还值得探讨.



图7 MDC 分类结果图

Fig.7 The classification result based on MDC

REFERENCES

- [1] Lothar Hemes, Dieter Frieau, Jan Puzicha, *et al* Support Vector Machines for Land Usage Classification in Landsat TM Imagery[J]. In: Proc. IGARSS '99, 1: 348—350.
- [2] Huang C, Davis L S, Townshend J R G. An assessment of support vector machines for land cover classification[J]. *International Journal of Remote Sensing*, 2002, 23: 725—749.
- [3] Martin Brown, Hugh G Lewis, Steve R Gunn. Linear Spectral Mixture Models and Support Vector Machines for Remote Sensing[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2000, 38(5): 2346—2360.
- [4] DONG Guang-Jun, ZHANG Yong-Sheng, FAN Yong-Hong. Image fusion for hyperspectral data of PHI and high-resolution aerial image[J]. *J. Infrared Millim. Waves* (董广军, 张永生, 范永弘. 基于多特征多分辨率融合的高光谱图像分类. *红外与毫米波学报*), 2006, 25(2): 123—126.
- [5] Zhang J, Zhang Y, Zhou T. Classification of hyperspectral data using support vector machine[C]. In: IEEE International Conference on Image Processing, 2001: 882—885.
- [6] ZHANG Xue-Gong. Introduction statistical learning theory and support vector machines[J]. *Acta Automatica Sinica* (张学工. 关于统计学习理论与支持向量机. *自动化学报*), 2000, 1(26): 32—42.
- [7] Vapnik V N. *Statistical Learning Theory*[M]. New York: Wiley, 1998.
- [8] Kreel U. *Pairwise Classification and Support Vector Machines*[M]. In *Advances in Kernel Methods: Support Vector Learning*, B. Scholkopf, C. J. C. Burges, and A. J. Smola (eds), The MIT Press, Cambridge, MA, 1999: 255—268.
- [9] Hsu C W, Lin C J. A comparison of methods for multi-class support vector machines[J]. *IEEE Transactions on Neural Networks*, 2002, 13(2): 415—425, 2002.
- [10] Boardman J W, Kruse F A. Automated spectral analysis: a geological example using AVIRIS data, north Grapevine Mountains[C]. Nevada: In Proceeding, ERM Tenth Thematic Conference on Geologic Remote Sensing, Environmental Research Institute of Michigan, Ann Arbor, MI, 1994, I407-I418.