一种改进的多类支持向量机超光谱图像分类方法

赵春晖,陈万海,万 建
(哈尔滨工程大学 信息与通信工程学院,黑龙江 哈尔滨 150001)

摘 要:支持向量机(SVM)是建立在统计学理论基础上的一种机器学习方法,用于解决二类分类问题,如何有效地将其推广到多类分类问题是一个正在研究的课题.总结了现有的主要的支持向量机多类分类算法,并在 1-a-1 SVM 分类算法基础上提出一种二次分类的方法.改良了惩罚因子,提高了不易分的类别之间的可分程度.通过对超光谱图像进行分类实验,结果表明该方法具有较高的分类精度.

关键词:支持向量机;二次分类;多类支持向量机
中图分类号:TN919.81 **文献标识码**:A **文章编号**:1673-4785(2008)01-0077-06

An improved hyperspectral image classification method for a multiclass support vector machine

ZHAO Chun-hui, CHEN Wan-hai, WAN jian
(College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China)

Abstract: SVM is a machine learning method developed on the basis of statistics theory and originally designed for binary classification problems. The most effective way to extend it for multiclass classification is still an area of considerable discussion. This paper presental a secondary classification method based on 1-a-1 SVM classification algorithm after a general overview of typical methods for a multiclass SVM. Our method improves the penalty factors, so it enhances the divisibility of classes that were difficult to classify. Experimental results of hyperspectral image classification showed that the suggested multiclass SVM has higher classification precision.

Key words: support vector machine; secondary classification; multiclass SVM

支持向量机(support vector machine, SVM)是 Vapnik 等在统计学习理论的基础上发展的一种新的模式识别方法,在解决小样本、非线性及高维模式识别问题中表现出许多特有的优势.最初支持向量机是用以解决两类分类问题,不能直接用于多类分类,而实际应用中遇到的多是多类分类问题.目前已经有许多算法将 SVM 推广到多类分类问题,这些算法统称为多类支持向量机.

1-a-1 SVM 分类方法是一个具有代表性的多类支持向量机算法,它在一定程度上改善了传统的多类支持向量机存在的错分、拒分区域,可以获得较好的分类效果.但当其应用到数据量大、维数高的超光谱图像时,就会遇到如何选择最佳惩罚因子和最优权向量系数的问题.由于目前尚无具体理论来指导最佳惩罚因子的选择,因此需要依赖大量的试验和研究者的经验,这对于高维的超光谱图像来说是很不现实的.文中针对这种情况提出了一种改进的多类支持向量机分类方法,即在 1-a-1 SVM 分类结果的基础上进行二次分类,以改善错分样本较多的类别之间的混淆程度.

1 支持向量机的分类原理

支持向量机的分类原理可概括为:寻找一个最优分类超平面,使得训练样本中的两类样本点能被无错误的分开,并且要使两类的分类间隔最大;而对线性不可分问题,通过核函数将低维输入空间的数

收稿日期:2007-06-06.
基金项目:高等学校博士学科点基金资助项目(20060217021);黑龙江省自然科学基金资助项目(ZJG0606-01).
通讯作者:赵春晖. E-mail:zhaochunhui@hrbeu.edu.cn.

据映射到高维空间,从而将原低维空间的线性不可分问题转化为高维空间上的线性可分问题,然后在这个新空间中求取最优分类面.

支持向量机分类的目标就是根据结构风险最小化原则,构造一个目标函数,寻找一个满足分类要求的最优超平面,即寻找最优线性判别函数. 设$\{x_i\} \in \mathbf{R}^D$为两类的样本数据,$y_i \in \{+1, -1\}$为相应的类别标号,$i = 1, 2, \cdots, n$. 如果 x_i 属于第 1 类,则 $y_i = 1$;如果 x_i 属于第 2 类,则 $y_i = -1$. 线性判别函数的一般形式为 $g(x) = w \cdot x + b$,相应的分类面为 $x \cdot x + b = 0$. 为了使待分样本尽可能好地分开,要求分类间隔(可表示为 $2/\|w\|$)最大,这相当于使 $\|w\|$ 最小. 寻找最优分类面可转化为求解数学形式的优化问题,通常可分为 3 种情况[1].

1) 线性可分问题.

针对线性可分情况,问题的目标函数的数学形式为

$$\min \frac{1}{2}\|w\|^2. \tag{1}$$

2) 线性不可分问题.

在处理线性不可分问题时,引入松弛变量 ε_i,$i = 1, 2, \cdots, n$ 和惩罚因子 C,对错分样本进行条件控制. 这样,式(1)可重新描述为

$$\min_{w,b,\varepsilon} J(w, \varepsilon) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\varepsilon_i,$$
$$\text{s.t.} \quad y_i[(w \cdot x_i) + b] - 1 + \varepsilon_i \geq 0,$$
$$\varepsilon_i \geq 0, i = 1, 2, \cdots, n, C > 0. \tag{2}$$

3) 非线性可分问题.

对于非线性情况,引入核函数 $\phi(x)$ 将原数据空间的非线性问题转化为高维特征空间的线性问题,即把 x_i 替换为 $\phi(x_i)$,其相应的内积 $x_i \cdot x_i$ 替换为 $K(x_i \cdot x_i) = \phi(x_i)^{\mathrm{T}}\phi(x_i)$,式(2)变为

$$\min_{w,b,\varepsilon} J(w, \varepsilon) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\varepsilon_i,$$
$$\text{s.t.} \quad y_i[(w \cdot \phi(x_i)) + b] - 1 + \varepsilon_i \geq 0,$$
$$\varepsilon_i \geq 0, i = 1, 2, \cdots, n, C > 0. \tag{3}$$

对于上述凸优化问题,可引入拉格朗日乘子 α_i $(i = 1, 2, \cdots, n)$,根据目标函数及约束条件建立 Lagrange 函数并将其转化为下面的对偶问题,即满足约束条件:

$$\sum_{i=1}^{n} y_i\alpha_i = 0, 0 \leq \alpha_i \leq C, i = 1, 2, \cdots, n.$$

对 α_i 求解下列函数的最大值:

$$Q(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i \cdot x_j). \tag{4}$$

式中:$\alpha = [\alpha_1\ \alpha_2 \cdots \alpha_n]^{\mathrm{T}}$. 设此矩阵方程的解为 α^*. 这是一个不等式约束下二次函数极值问题,存在唯一解. 且根据 KKT 条件,这个优化问题的解必须满足:

$$\alpha_i(y_i[(w \cdot \phi(x_i)) + b] - 1 + \varepsilon_i) = 0,$$
$$i = 1, 2, \cdots, n.$$

因此,对多数样本的 α_i^* 将为 0,对分类问题不起什么作用,只有取值不为 0 的 α_i^* 对应的样本才决定最终分类结果,这样的样本称之为支持向量,它们通常只是全体样本中的很少一部分.

求解上述问题后得到最终的判别函数为

$$f(x) = \operatorname{sign}\{\sum_{i=1}^{n}\alpha_i^* y_i K(x_i, x) + b^*\}. \tag{5}$$

根据 $f(x)$ 的结果来确定 x 属于哪一类.

2 现有的多类支持向量机算法

目前对于多类分类问题,支持向量机的解决途径有 2 种[2-3]:一种是通过构造多个两类 SVM 分类器并将它们组合起来实现多类分类,将多类问题转化为两类分类问题;第 2 种是一次性解决多类分类问题,即把所有子分类器的参数直接放在一个最优化方程里同时优化,这种思想尽管看起来简洁,但在最优化问题求解过程中的变量远远多于第 1 种,训练速度及分类精度也不占优势,因此目前多采用第 1 种方法,文中介绍的这几种多类支持向量机都属于第 1 种方法.

2.1 1-a-r SVM 算法

1-a-r (one-against-rest) SVM 算法是解决多类分类问题的最早的方法,对于 $k(k \geq 2)$ 类 SVM 分类问题,把其中一类作为第 1 类,其余类视为另一类,自然地将 k 分类问题转化为 k 个两类分类问题,得到 k 个两类分类器,分类时未知样本最后的输出是两类分类器输出为最大的那一类.

2.2 1-a-1 SVM 算法

1-a-1 (one-against-one) SVM 算法是在每两类之间训练一个分类器,因此对于一个 k 类问题,训练阶段共构造 C_k^2 个两类分类器,每个分类器是取任意 2 个类别的数据进行训练. 在测试阶段可以采用不同的方式测试样本属于哪一类,最常用的一种方法是“最大投票法”,即每个两类分类器都对样本的类别进行判断,采用投票机制为其相应的类别“投上一票”,最后得票最多的类别即是该未知样本的所属类. 因此,文中在此算法的基础上,提出了一种改进的多类支持向量机分类方法.

3 二次分类的多类支持向量机

支持向量机应用于实际植被的模式识别时,常

会出现某两类或某几类植被的光谱特性极为相似的情况,它们在分类时往往会产生严重混淆的现象,导致整体分类效果不够理想. 这种现象产生的原因是由于植物种类的分类精度很大程度上取决于植物长势和叶面叶绿素含量,当不同物种叶面叶绿素的含量相近时,容易错分,这就取决于实际拍摄图片的时间和地点等各种因素. 针对此现象文中提出了一种二次分类的多类支持向量机,它是在 1-a-1 SVM 算法的基础上改进的多类支持向量机分类方法.

采用 1-a-1 SVM 进行分类时,训练阶段需要在每两类之间训练一个分类器,对于一个 k 类问题需要构造 C_k^2 个两类分类器. 对于第 i 类和第 j 类之间的训练,需要解决下列两类分类问题:

$$\min_{w^{ij},b^{ij},\varepsilon^{ij}} \frac{1}{2}(w^{ij})^{\mathrm{T}}w^{ij} + C\sum_t \varepsilon_t^{ij}(w^{ij})^{\mathrm{T}},$$

$$(w^{ij})^{\mathrm{T}}\phi(x_t) + b^{ij} \geq 1 - \varepsilon_t^{ij}, \text{If } y_t = i,$$

$$(w^{ij})^{\mathrm{T}}\phi(x_t) + b^{ij} \leq -1 + \varepsilon_t^{ij}, \text{If } y_t = j,$$

$$\varepsilon_t^{ij} \geq 0. \tag{6}$$

在测试时采用最大投票法,判断符号函数:

$$f(x) = \mathrm{sign}((w^{ij})^{\mathrm{T}}\phi(x) + b^{ij}). \tag{7}$$

若未知样本 x 属于第 i 类则第 i 类的票数加一,反之第 j 类加一,x 属于最后票数最多的那一类.

两类分类的支持向量机在训练阶段需要寻找最佳惩罚因子 C. 如果惩罚因子选择的不恰当,分类精度将会受到不良影响. 采用 1-a-1 SVM 算法进行分类时,若类别数 k 较大时,需要构造大量的两类分类器. 在实际应用中不可能做到把每个两类分类问题都做多次实验来寻找最佳惩罚因子,因此提出了二次分类的方法,分析采用 1-a-1 SVM 进行第 1 次分类的结果,找出混淆最严重的组合,对它们进行二次分类,寻找最佳惩罚因子重新分类,这样就可以减小这些类别之间的混淆程度. 二次分类算法在训练阶段的模型如图 1 所示.

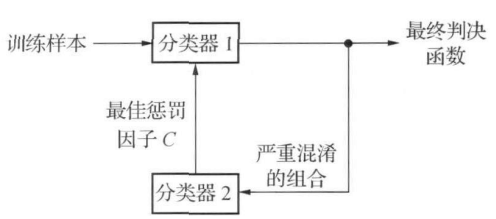


图 1　二次分类的训练阶段模型

Fig. 1　Training stage model of secondary classification

进行二次分类时需要单独解决这个组合的两类分类问题. 对于 k 类问题,假设有一对混淆严重的组合为第 i 类和第 j 类,其相应的对偶问题为

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

$$0 \leq \alpha_i \leq C, \sum_i^n y_i\alpha_i = 0, i = 1,2,\cdots,n. \tag{8}$$

式中:C 为惩罚因子,用来对错分样本进行条件控制,根据经验和多次实验得到最佳惩罚因子 C^*,设此时的拉格朗日乘子 α^* 为不等式约束的最优解,由此得到权向量系数为

$$w^* = \sum_i^n \alpha_i^* y_i x_i. \tag{9}$$

把这个最佳惩罚因子 C^* 和最优的权系数向量 x^* 反馈到 1-a-1 SVM 算法中重新分类.

4　实验结果与分析

为了验证所提方法的有效性,采用 AVIRIS 超光谱遥感图像进行实验. 该图像取自 1992 年 6 月拍摄的美国印第安纳州西北部印第安遥感试验区的一部分[4],它包含了农作物和森林植被的混合区.

文中实验分析共有 3 组分类实验,利用自适应波段选择的方法[5]从原始图像的 220 个波段中选取 3 个波段图像:9 波段、18 波段和 50 波段作为研究对象,对照真实地物图,进行玉米、大豆、干草、林地、牧场和草地这 6 类地物的分类实验.

文中采用混淆矩阵和总体分类精度来分析分类效果,混淆矩阵定义如下[6]:

$$M = \begin{bmatrix} m_{11} & \cdots & m_{1n} \\ \cdots & \ddots & \cdots \\ m_{n1} & \cdots & m_{nn} \end{bmatrix}. \tag{10}$$

式中:m_{ij} 表示实验区内应属于第 i 类的样本被分到第 j 类中去的样本总数,n 为类别数. 如果混淆矩阵中对角线上的元素值愈大,则表示分类结果的可靠性愈高,如果混淆矩阵中非对角线上的元素值愈大,则表示错误分类的现象愈严重,文中实验结果的混淆矩阵采用表格的形式进行描述. 像素的总体分类精度(OA)的定义如下:

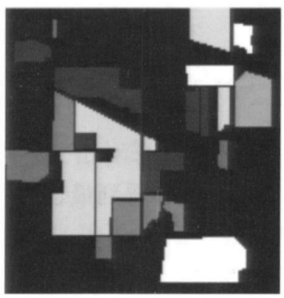
$$OA = \frac{\sum_i^n m_{ii}}{N}. \tag{11}$$

式中:N 为参与分类的总样本数,m_{ii} 为第 i 类正确分类的样本数.

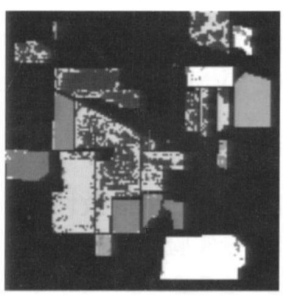
将 1-a-1 SVM 和改进的二次分类算法应用到 3 组波段的地物分类中,并从灰度图和混淆矩阵 2 个方面来比较它们的分类效果.

1) 9 波段的地物分类.

基于 1-a-1 SVM 的分类灰度图和混淆矩阵如图 2、表 1 所示.



(a)真实地物灰度图



(b)9 波段分类灰度图

图 2 基于 1-a-1 SVM 的分类灰度图

Fig. 2 Grey image by 1-a-1 SVM classification

表 1 9 波段的混淆矩阵

Table 1 Mixture matrix of 9-waveband

	玉米	牧场	草地	干草	大豆	林地
玉米	985	0	4	1	444	0
牧场	7	315	43	93	30	9
草地	0	2	722	5	17	1
干草	0	0	2	486	1	0
大豆	1 027	0	10	2	1 429	0
林地	0	61	15	0	7	1 211

总体分类精度:74.30 %.

从混淆矩阵中可以看出玉米和大豆这两类混淆最严重,因此采用文中提出的二次分类方法,训练出玉米和大豆两类别之间的最佳惩罚因子,并重新对地物图像进行分类.

经过多次实验,得到玉米和大豆的惩罚因子 C 和分类精度之间的关系如表 2 所示.

表 2 9 波段的惩罚因子和分类精度关系

Table 2 Relation between penalty factor and classification precision for 9-waveband

惩罚因子 C	∞	2000	201	196～200	195	100
分类精度/ %	62.10	63.92	64.17	64.20	64.17	63.86

从表 5 中选取精度最大时对应的 C 为 200.

二次分类的分类灰度图和混淆矩阵如图 3、表 3 所示.

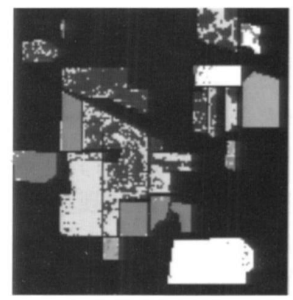


图 3 基于二次分类的 9 波段分类灰度图

Fig. 3 9-waveband grey image by secondary classification

表 3 9 波段的混淆矩阵

Table 3 Mixture matrix of 9-waveband

	玉米	牧场	草地	干草	大豆	林地
玉米	1 072	0	4	1	357	0
牧场	7	315	43	93	30	9
草地	0	2	722	5	17	1
干草	0	0	2	486	1	0
大豆	1 019	0	10	2	1 437	0
林地	0	61	15	0	7	1 211

总体分类精度:74.9 %.

2) 18 波段的地物分类.

基于 1-a-1 SVM 的分类灰度图和混淆矩阵如图 4、表 4 所示.

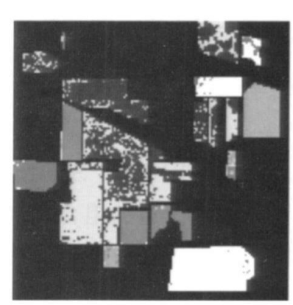


图 4 基于 1-a-1 SVM 的 18 波段分类灰度图

Fig. 4 18-waveband grey image by 1-a-1 SVM classification

表 4 18 波段的混淆矩阵

Table 4 Mixture matrix of 18-waveband

	玉米	牧场	草地	干草	大豆	林地
玉米	1 030	0	4	3	397	0
牧场	7	319	42	98	26	5
草地	0	2	721	7	16	1
干草	0	0	2	486	1	0
大豆	1 069	0	8	4	1 387	0
林地	0	58	15	0	6	1 215

总体分类精度:74.44 %.

18 波段的玉米和大豆的惩罚因子 C 和分类精

度之间的关系如表 5 所示.

表 5　18 波段的惩罚因子和分类精度关系

Table 5　Relation between penalty factor and classification precision for 18-waveband

惩罚因子 C	∞	2000	94	89～93	88	85
分类精度/%	62.17	62.17	62.97	62.99	62.97	62.92

从表 5 中选取 18 波段的最佳 C 为 90.

二次分类的分类灰度图和混淆矩阵如图 5、表 6 所示.

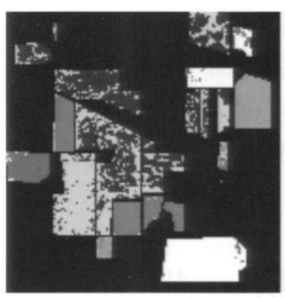


图 5　基于二次分类的 18 波段分类灰度图

Fig. 5　18-waveband grey image by secondary classification

表 6　18 波段的混淆矩阵

Table 6　Mixture matrix of 18-waveband

	玉米	牧场	草地	干草	大豆	林地
玉米	1 046	0	4	3	381	0
牧场	7	319	42	98	26	5
草地	0	2	721	7	16	1
干草	0	0	2	486	1	0
大豆	1 053	0	8	4	1 403	0
林地	0	58	15	0	6	1 215

总体分类精度:75.67 %.

3) 50 波段的地物分类.

基于 1-a-1 SVM 的分类灰度图和混淆矩阵如图 6、表 7 所示.

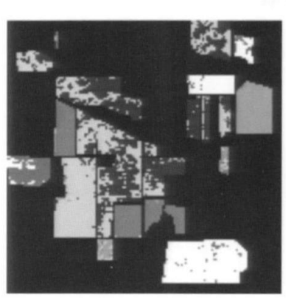


图 6　基于 1-a-1 SVM 的 50 波段分类灰度图

Fig. 6　50-waveband grey image by 1-a-1 SVM classification

表 7　50 波段的混淆矩阵

Table 7　Mixture matrix of 50-waveband

	玉米	牧场	草地	干草	大豆	林地
玉米	952	0	4	9	469	0
牧场	6	266	27	92	46	60
草地	0	0	725	18	3	1
干草	0	0	3	486	0	0
大豆	878	0	13	17	1 560	0
林地	0	77	12	0	5	1 200

总体分类精度:74.89 %.

50 波段的玉米和大豆的惩罚因子 C 和分类精度之间的关系如表 8 所示.

表 8　50 波段的惩罚因子和分类精度关系

Table 8　Relation between penalty factor and classification precision for 50-waveband

惩罚因子 C	∞	500	56	43～55	42	0
分类精度/%	69.95	70.26	70.85	70.89	70.85	65.22

从表中选取 50 波段的最佳 C 是 50.

二次分类的分类灰度图和混淆矩阵如图 7、表 9 所示.

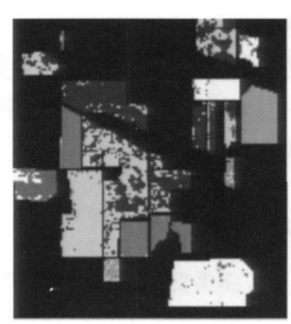


图 7　基于二次分类的 50 波段分类灰度图

Fig. 7　50-waveband grey image by secondary classification

表 9　50 波段的混淆矩阵

Table 9　Mixture matrix of 50-waveband

	玉米	牧场	草地	干草	大豆	林地
玉米	1 126	0	4	9	295	0
牧场	9	266	27	92	43	60
草地	0	0	725	18	3	1
干草	0	0	3	486	0	0
大豆	828	0	13	17	1 610	0
林地	0	77	12	0	5	1 200

总体分类精度:78.12 %.

表 3、6、9 所示的是利用文中提出的改进的二次分类方法的结果,将其与 1-a-1 SVM 分类方法的混

淆矩阵表1、4、7相比较会发现,之前混淆较严重的玉米—大豆这两类在新方法的分类结果中混淆样本有了很大的改观.从分类的总体精度上还可以看出,选取的波段数与分类精度也有很大的关系,波段数目越多,所包含的光谱特性就越丰富,得到的分类结果就越好.

分类总体精度提高不上去的主要原因是原始图像的精度不高,由于使用的AVIRIS图像是从高空拍摄,其分辨率仅为20 m×20 m,像元混合的概率很大,而且图片拍摄于6月,玉米和大豆正处于生长的早期,它们所反应的光谱特性极其相似,在这些客观原因的影响下,这两类的分类精度不可能大幅度的提高,仍存在一定的分类误差.

5 结束语

该文提出的二次分类方法保留了1-a-1 SVM分类的优点,改善了超光谱图像分类中某些植被由于光谱特性相似而产生严重混淆的问题.利用二次分类方法可以在短时间内取得混淆最严重的类别的最佳惩罚因子,将其应用到1-a-1 SVM的训练过程中,得到最优的权系数向量和支持向量,改善了分类精度.它弥补了1-a-1 SVM算法不能确定最佳惩罚因子的缺陷,提高了支持向量机应用到超光谱图像中的分类效果.

在许多的实际应用中都需要解决多类别的分类问题,如何有效地将支持向量机推广到多类分类问题仍有广阔的研究空间.

参考文献:

[1]任建峰,郭 雷.多类支持向量机的自然图像分类[J].西北工业大学学报,2005,23(3):295-298.
REN Jianfeng, GUO Lei. Improving scene image classification with multi-class SVMs[J]. Journal of Northwestern Polytechnical University, 2005, 23(3): 295-298.

[2]HSU C W, LIN C J. A comparison of methods for multiclass support vector machines[J]. IEEE Transactions on Neural Networks, 2002, 13(2): 415-425.

[3]刘志刚,李德仁,秦前清,等.支持向量机在多类分类问题中的推广[J].计算机工程与应用,2004(7):10-13.
LIU Zhigang, LI Deren, QIN Qianqing, et al. An analytical overview of methods for multi-category support vector machines[J]. Computer Engineering and Applications, 2004(7): 10-13.

[4]LANDGREBE D. Multispectral data analysis: a signal theory perspective[R]. West Lafayette, USA: University, 1998.

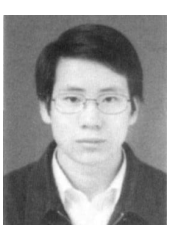
[5]刘春红,赵春晖,张凌雁.一种新的高光谱遥感图像降维方法[J].中国图像图形学报,2005,10(3):218-222.
LIU Chunhong, ZHAO Chunhui, ZHANG Lingyan. A new method of hyperspectral remote sensing image dimensional reduction[J]. Journal of Image and Graphic, 2005, 10(3): 218-222.

[6]赵英时.遥感应用分析原理与方法[M].北京:科学出版社,2003.

作者简介:

赵春晖,男,1965年生,教授,博士生导师.获省部级科技奖5项.主要研究方向为智能信息处理技术、图像处理.出版著作3部,发表论文200余篇.

陈万海,男,1963年生,副教授,博士研究生.主要研究方向为超光谱遥感图像处理技术,发表论文18篇.



万 建,男,1980年生,博士研究生.主要研究方向为信号与图像处理,发表论文5篇.