利用 SVM 的极化 SAR 图像特征选择与分类

吴永辉　　计科峰　　李　禹　　郁文贤
(国防科学技术大学电子科学与工程学院　长沙　410073)

摘　要：该文提出一种新的利用 SVM 的特征选择算法，并将其融入到极化 SAR 图像分类过程中，构成一种新的基于 SVM 的分类方法。其中，特征选择算法利用支持向量个数作为特征评估指标，并以顺序后退法作为搜索策略。真实数据的实验结果表明，该分类方法能有效降低 SVM 分类器对自身参数的敏感性，与利用原始特征集和经典的 RELIEF-F 的分类方法相比，该方法能以更少(或相当)的特征个数，在更广泛的 SVM 参数取值范围内获得更高的分类精度。

关键词：合成孔径雷达(SAR)；雷达极化；特征选择；分类；支持向量机(SVM)
中图分类号：TP753　　**文献标识码**：A　　**文章编号**：1009-5896(2008)10-2347-05

Feature Selection and Classification of Polarimetric SAR Images Using SVM

Wu Yong-hui　　Ji Ke-feng　　Li Yu　　Yu Wen-xian
(*School of Electronics Science and Engineering*, *National University of Defense Technology*, *Changsha* 410073, *China*)

Abstract: A new feature selection algorithm is presented using SVM, and then it is integrated into the classification procedure of polarimetric SAR images to construct a novel SVM-based classification method. In the novel method, the sequential backward selection strategy is used to search feature subsets, and the number of support vectors is taken as the estimation index. Compared with those using the initial feature set and the classical RELIEF-F algorithm, higher classification accuracy with less or equivalent number of features is observed in a wider range of SVM parameters using the novel method.

Key words: Synthetic Aperture Radar (SAR); Radar polarimetry; Feature selection; Classification; Support Vector Machine (SVM)

1　引言

极化 SAR 测量得到多个极化通道数据，能更完整地表征地物散射。如何利用这些信息提高分类精度是极化 SAR 图像分类的热点问题。解决该问题有两种途径：一种是引入新方法或改进已有方法，更充分地利用分类信息[1]；另一种是提取与地物类别相关性更强的新特征，提高特征集本身区分地物的能力[2]。本文从第一种途径入手，将支持向量机(SVM)[3]用于极化 SAR 图像分类，重点研究了特征选择及其在分类过程中的作用。SVM 适用于小样本情况，具有良好的泛化性[3]。然而，SVM 未利用数据统计知识，因此构造合适的特征集非常重要。文献[1, 4]通常依据经验选取特征，不仅降低了自动化程度，同时，对于不同场景仅凭经验难以得到合适特征集，这将直接影响分类精度[5]。另外，虽然 SVM 通过将原始特征空间数据映射到高维空间以获得较好的性能，但原始特征集仍对分类结果有很大影响。特征选择通过选出满足给定准则的特征集改善分类结果，能较好地解决这两个问题。

有鉴于此，本文提出一种新的利用 SVM 的特征选择算法，并将其用作分类的预处理，构成一种基于 SVM 的极化 SAR 图像分类方法。该分类方法包括两步：第 1 步，以 SVM 输出的支持向量个数作为评估特征优劣的指标，采用顺序后退法(Sequential Backward Selection, SBS)[6]搜索特征子集，得到结果特征集；第 2 步，将结果特征集输入 SVM 分类器，得到最终的分类结果。

2007-03-12 收到，2007-11-12 改回

2　SVM 原理简介

假定原始特征空间中的训练数据 $\{(\boldsymbol{x}_1,y_1),\cdots,(\boldsymbol{x}_l,y_l)\}$ $(\boldsymbol{x}_i\in R^n,\ y_i\in\{-1,1\},\ i=1,2,\cdots,l)$ 可被超平面 $\langle\boldsymbol{w},\boldsymbol{x}\rangle+b=0(b\in R)$ 线性划分为两类，超平面必定满足 $y_i(\langle\boldsymbol{w},\boldsymbol{x}_i\rangle+b)\ge 1\ (i=1,2,\cdots,l)$，其中 $\langle\cdot,\cdot\rangle$ 表示矢量内积。SVM 所求的最优超平面要使得分类间隔最大，这等价于求解二次规划问题：

$$\min_{\boldsymbol{w},b}\Phi(\boldsymbol{w})=\langle\boldsymbol{w},\boldsymbol{w}\rangle/2 \tag{1}$$

$$\text{s.t. } y_i\left(\langle\boldsymbol{w},\boldsymbol{x}_i\rangle+b\right)\ge 1,\quad i=1,2,\cdots,l \tag{2}$$

引入 Lagrange 算子 $\alpha_i\ge 0$，可得到上述问题的唯一解 $\boldsymbol{\alpha}_i^*=[\alpha_1^*,\alpha_2^*,\cdots,\alpha_l^*]^{\mathrm{T}}$，满足

$$\alpha_i^*\left[y_i\left(\langle\boldsymbol{w}^*,\boldsymbol{x}_i\rangle+b^*\right)-1\right]=0,\quad i=1,2,\cdots,l \tag{3}$$

基于最优分类超平面的判决函数为

$$f(\boldsymbol{x})=\mathrm{sgn}\left(\langle \boldsymbol{w}^*,\boldsymbol{x}\rangle+b^*\right)=\mathrm{sgn}\left(\sum\nolimits_{i\in \mathrm{SV}} y_i\alpha_i^*\langle \boldsymbol{x}_i,\boldsymbol{x}\rangle+b^*\right) \quad (4)$$

若原始特征空间中训练数据线性不可分，可引入松弛变量 $\xi_i(\geq 0)$ 和惩罚因子 $C(>0)$ 计算判决函数。

3 极化 SAR 图像分类特征

本文选取较为常用的 15 个特征构成全极化 SAR 图像分类的原始输入特征集：

$$F=\{C_{11},C_{22},C_{33},|C_{12}|,|C_{13}|,|C_{23}|,\varphi(C_{12}),\varphi(C_{13}),\varphi(C_{23}),\ \mathrm{span},\lambda_1,\lambda_2,\lambda_3,H,\alpha\} \quad (5)$$

式中 C_{ij} 为协方差矩阵第 i 行第 j 列元素，$|\cdot|$ 表示取模，$\varphi(\cdot)$ 表示取幅角，span 为总功率图，$\lambda_i(i=1,2,3)$ 为相干矩阵特征值，H 和 α 分别为目标散射极化熵和表征目标散射机理的角度[2]。

对于双极化数据，选用以下 9 个常用特征构成双极化 SAR 图像分类的原始输入特征集：

$$F=\{C_{11},C_{22},|C_{12}|,\varphi(C_{12}),\mathrm{span},\lambda_1,\lambda_2,H,\alpha\} \quad (6)$$

4 利用 SVM 的极化 SAR 图像特征选择与分类

式(3)表明，若样本到分类超平面的距离是最短距离，即 $y_i(\langle \boldsymbol{w}^*,\boldsymbol{x}_i\rangle+b^*)=1$，则 $\alpha_i^*\neq 0$，对应的样本称为支持向量(Support Vector, SV)。由式(4)可以看出，选出任意一个非支持向量，利用训练集的剩余样本均可对其正确分类。训练集误分概率 P_e 满足[3]：

$$E\{P_e\}\leq N_{\mathrm{SV}}/l \quad (7)$$

式中 N_{SV} 为支持向量个数。易知，支持向量越少，分类精度越高，SVM 泛化能力越强。对于参数确定的 SVM，支持向量越少，表明各类样本可分性越好，分类精度就越高。图 1 为样本可分性与支持向量个数关系示意图。容易看出，对分类性能好的特征，各类样本可分性好， SVM 构造分类超平面时所需支持向量较少，如图 1(a)所示；而对于分类性能较差的特征，各类样本交界处分布情况复杂，SVM 需要较多支持向量才能构造分类超平面，如图 1(b)所示。

由此，本文用支持向量个数作为评估特征优劣的指标，特征评估准则为

$$J(F_{\mathrm{out}})=\mathrm{NSV}(F_{\mathrm{out}})=\min_{A\subseteq F_{\mathrm{in}}}J(A)=\min_{A\subseteq F_{\mathrm{in}}}\mathrm{NSV}(A) \quad (8)$$

式中 F_{out} 为输出的 d 维特征集，F_{in} 为输入的 k 维特征集，A 为 F_{in} 的子集， NSV(·) 表示求特征(或特征集)中各类地物的平均支持向量个数。式(8)表明，特征子集的支持向量越少越好。

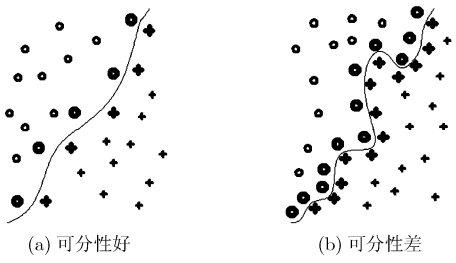

图 1 样本可分性与 N_{SV} 关系

将式(8)作为特征评估准则，采用 SBS 作搜索策略，本文提出一种新的特征选择算法，称为基于支持向量个数的后向选择法(Number of Support Vectors based Backward Selection, NSVBS)。k 个特征构成的特征集包含 2^k-1 个子集，用穷举法搜索全局最优解是不现实的。由于支持向量个数不随特征集单调变化，即对于 $F_1\subset F_2$，不能保证 $J(F_1)\geq J(F_2)$，故无法用分支定界法[7]获得最优子集。而 SBS 作为常用次优搜索算法，能极大减少计算量，并保证结果特征集具有较好分类性能[6]。NSVBS 具体步骤为

(1)输入包含 k 个特征的原始特征集 $F=\{\boldsymbol{f}_i,i=1,2,\cdots,k\}$；

(2)选取训练数据，计算各特征 $\boldsymbol{f}_i(i=1,2,\cdots,k)$ 的平均支持向量个数 $\bar{N}_{\mathrm{SV}}^{(i)}=\mathrm{NSV}(\boldsymbol{f}_i)$；

(3)按照 $\bar{N}_{\mathrm{SV}}^{(i)}$ 的降幂对 F 中所有特征排序，得到排序特征集 $F'=\{\boldsymbol{f}_i',i=1,2,\cdots,k\}$，其中 $\mathrm{NSV}(\boldsymbol{f}_i')\geq\mathrm{NSV}(\boldsymbol{f}_j'),(i<j)$；

(4)按 $\bar{N}_{\mathrm{SV}}^{(i)}$ 的降幂逐个选择 F' 中的特征 $\boldsymbol{f}_i'(i=1,2,\cdots,k)$，将其作为待剔除特征；

(5)对于第 n 个待剔除特征 $\boldsymbol{f}_n'$，若 $\mathrm{NSV}(F'\setminus\boldsymbol{f}_n')<\mathrm{NSV}(F')$，则令 $F'=F'\setminus\boldsymbol{f}_n'$，即，从 F' 中剔除 $\boldsymbol{f}_n'$，否则，保留 $\boldsymbol{f}_n'$；

(6)令计数器 $n=n+1$，若 $n\leq k$，返回步骤(4)；否则，特征选择结束，得到包含 d 个已选特征的结果特征集 $F''=\{\boldsymbol{f}_i'',i=1,2,\cdots,d\}=F'$。

图 2 给出了本文基于 SVM 的极化 SAR 图像分类方法的简要流程。

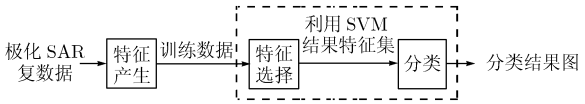

图 2 基于 SVM 的极化 SAR 图像分类方法流程图

5 实验结果及分析

5.1 Flevoland 全极化数据特征选择与分类

图 3 给出了 NASA/JPL 的 AIRSAR 获取的 L 波段荷兰 Flevoland 全极化 4 视数据，大小为 400×300 像素，包含 8 类地物。总功率图经 3×3 矩形窗滤波，图 3(c)中各类训练数据大小均为 20×17 像素。为验证本文中式(8)作为特征评估准则的合理性，从式(5)所示 F 中随机选取 $d(=1,3,5,7,9,11,13,15)$ 个特征构成特征集(共 71 个不同的分类特征集，d 的前 7 个值各取 10 个子集)，用径向基核函数 SVM(形状参数 $\sigma=1$，惩罚因子 $C=100$)得到不同特征个数时分类精度与支持向量个数的关系如图 4 所示。从图 4 可以看出，分

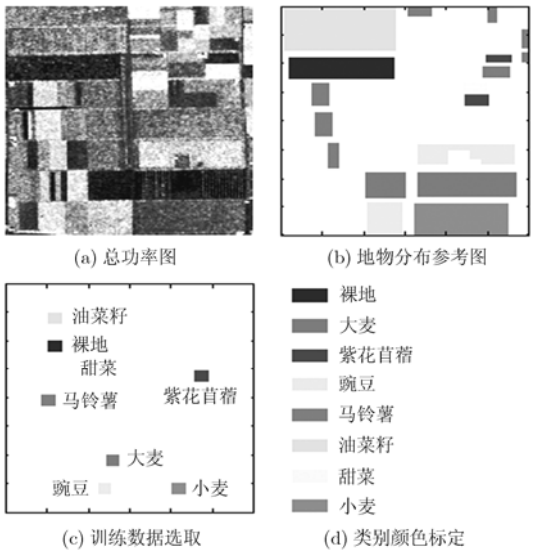

图 3 荷兰 Flevoland 地区 L 波段全极化数据

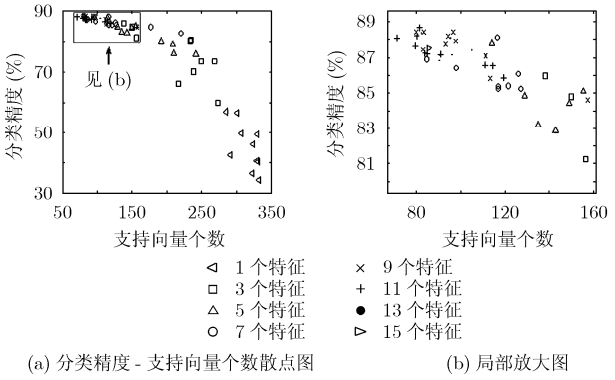

图 4 不同特征个数时分类精度与支持向量个数的关系

类精度随支持向量个数减少而提高的整体趋势非常明显，表明特征集所需支持向量个数越少，则不同类别样本的可分性越好，分类精度越高。图 4 表明利用式(8)作为评估准则是合理的。

利用径向基核函数 SVM 对式(5)所示 F (经 3×3 矩形窗滤波以进一步减弱相干斑影响)进行特征选择，并将 F'' 输入 SVM 进行分类，得到 $\sigma=0.1,0.3,0.5,0.75,1$ 时分类精度随 C 变化的曲线如图 5。图 5 同时给出了以 SVM 为分类器，对原始特征集 F 和 RELIEF-F[8]特征选择结果进行分类的精度。图例中 $w_{\text{th}}=0.015,\ 0.019$ 为 RELIEF-F 的门限。表 1 为对应的结果特征集中特征个数。

从图 5 和表 1 可知，利用 RELIEF-F 虽能一定程度减少特征个数，但精度比原始特征集略有下降，NSVBS 特征个数少于 RELIEF-F 或相当，精度却明显高于利用 RELIEF-F 和原始特征集的方法。$\sigma=0.1$ 时，用 RELIEF-F 和原始特征集的方法与本文方法精度相差最大，后者精度约为 88%，比 RELIEF-F 和原始特征集高 14%和 13%以上。随着 σ 增大，前两者精度虽有所提高，但仍低于本文方法。由图 5 可知，

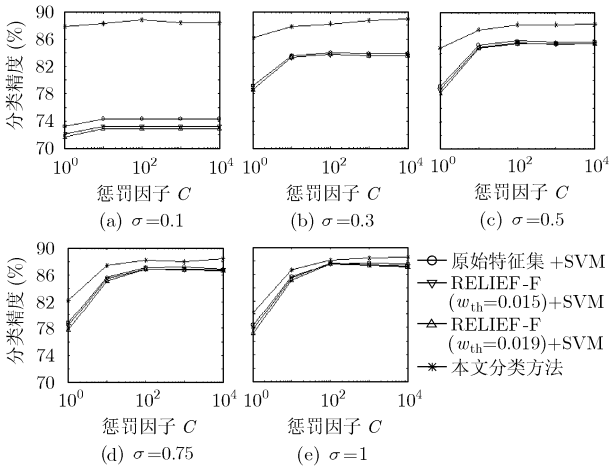

图 5 不同 SVM 参数时本文分类方法与其它方法分类精度比较

表 1 NSVBS 与 RELIEF-F 的结果特征集中特征个数比较

特征选择算法		结果特征集中特征个数				
		$C=1$	$C=10$	$C=10^2$	$C=10^3$	$C=10^4$
NSVBS	$\sigma=0.1$	9	7	7	8	5
	$\sigma=0.3$	12	11	8	9	7
	$\sigma=0.5$	12	12	12	8	7
	$\sigma=0.75$	12	12	11	10	7
	$\sigma=1$	12	12	12	10	11
RELIEF-F	$w_{\text{th}}=0.015$			12		
	$w_{\text{th}}=0.019$			11		

σ 变化时，用 RELIEF-F 和原始特征集得到的精度波动很大，前者最大波动幅度大于 14%($C=100$ 时)，后者大于 13%($C=1000$ 时)，两者最低精度小于 74%；而本文方法最大波动幅度低于 7.5%($C=1$ 时)，约为前两者的一半，同时能始终保持较高精度。这说明利用 RELIEF-F 和原始特征集的分类结果受 SVM 参数影响大，而 NSVBS 能在广泛的 SVM 参数范围内获得合适的分类特征集，减弱了 SVM 参数对分类精度的影响，使 SVM 分类性能更加稳定，增强了 SVM 分类器的自适应性。

令 $C_i=1,10,10^2,10^3,10^4,(i=1,\cdots,5)$，式(5)所示原始特征集 F 为全集。给定 σ，则 NSVBS 的结果特征集可表示为 F''_{C_i}。表 2 给出了 C 取不同值时，$F''_{C_i}(i=1,\cdots,5)$ 都选中与都未选中的特征，即 $\bigcap_{i=1}^{5}F''$ 与 $\bigcap_{i=1}^{5}\overline{F''_{C_i}}$。作为对比，表 2 同时列出了 RELIEF-F 选中与未选中的特征。

表 2 中，HV 极化的幅度 $|C_{12}|$ 是唯一被 NSVBS 和 RELIEF-F 都选中的特征，表明了该特征在农作物分类中的重要性。λ_3，H 和 α 每次都不被 NSVBS 选中，但被 RELIEF-F 选中，而利用 NSVBS 的结果特征集得到的分类精度始终高于利用 RELIEF-F 的结果，说明这 3 个特征不适合用于作物的精细分类。

表 2 惩罚因子 C =1, 10, 10^2, 10^3, 10^4 时，NSVBS 与 RELIEF-F 的结果特征集中选中和未选中特征列表

特征选择算法		选中特征($\bigcap_{i=1}^{5} F_{C_i}$)	未选中特征($\bigcap_{i=1}^{5} \overline{F_{C_i}}$)
NSVBS	σ=0.1	$\|C_{12}\|$, $\|C_{23}\|$, $\varphi(C_{13})$, $\varphi(C_{23})$	C_{11}, span, λ_2, λ_3, H, α
	σ=0.3	$\|C_{12}\|$, $\|C_{13}\|$, $\varphi(C_{12})$, $\varphi(C_{13})$, $\varphi(C_{23})$, λ_1	λ_3, H, α
	σ=0.5	C_{22}, $\|C_{12}\|$, $\|C_{13}\|$, $\|C_{23}\|$, $\varphi(C_{12})$	λ_3, H, α
	σ=0.75	C_{22}, $\|C_{12}\|$, $\|C_{13}\|$, $\|C_{23}\|$, $\varphi(C_{12})$, $\varphi(C_{13})$, $\varphi(C_{23})$	λ_3, H, α
	σ=1	C_{11}, C_{22}, C_{33}, $\|C_{12}\|$, $\|C_{13}\|$, $\|C_{23}\|$, $\varphi(C_{12})$, $\varphi(C_{13})$, λ_2	λ_3, H, α
RELIEF-F	w_{th}=0.015	C_{11}, C_{22}, C_{33}, $\|C_{12}\|$, $\|C_{13}\|$, $\|C_{23}\|$, $\varphi(C_{12})$, $\varphi(C_{23})$, λ_1, λ_3, H, α	$\varphi(C_{13})$, span, λ_2
	w_{th}=0.019	C_{11}, C_{22}, C_{33}, $\|C_{12}\|$, $\|C_{13}\|$, $\|C_{23}\|$, $\varphi(C_{12})$, $\varphi(C_{23})$, λ_3, H, α	$\varphi(C_{13})$, span, λ_1, λ_2

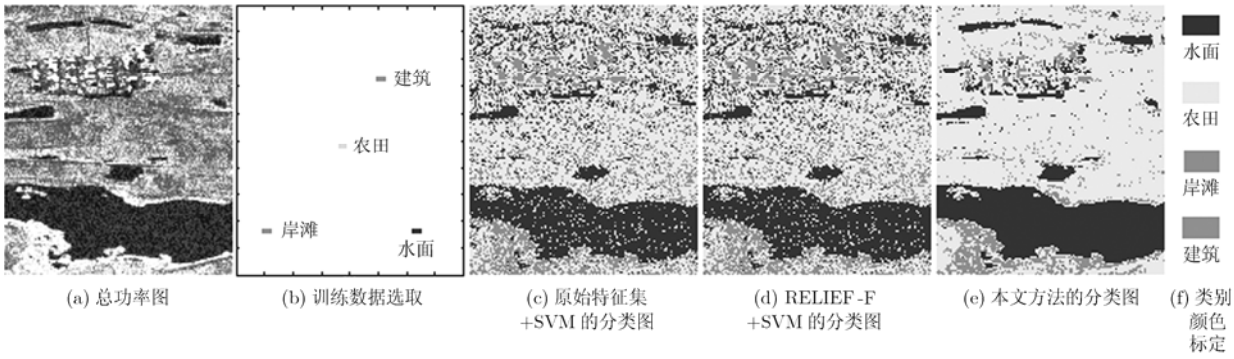

图 6 国内 HH-HV 双极化数据

5.2 国内某地区双极化数据特征选择与分类

实验 2 用到的是国内某地区 HH-HV 双极化 8 视 SAR 数据，大小为 500×400 像素，大致包括水面、农田、岸滩和建筑 4 类，如图 6(a)。仍选用径向基核函数 SVM 进行特征选择和分类，原始特征集如式(6)，特征集经 3×3 矩形窗滤波。训练数据根据总功率图选取，均为 9×17 像素。由于缺少精确地物分布图，无法计算分类精度，故采用目视评测。图 6 给出了 $\sigma=0.1$ 和 $C=100$ 时用原始特征集，RELIEF-F ($w_{\text{th}}=0.0065$)和本文方法得到的分类图。大量实验表明，$\sigma=0.1,0.3,0.5,0.75,1$ 和 $C=1,10,10^2,10^3,10^4$ 时，实验结果与图 6 类似，故略去。

比较图 6(c)和图 6(a)可见，原始特征集的分类图中，部分水面像素误分为农田和岸滩，许多农田像素误分为其它 3 类，分类图模糊不清。NSVBS 的结果特征集 $F''=\{C_{11},C_{22},|C_{12}|,\text{span},\lambda_1,\lambda_2\}$ 包含 6 个特征，与 RELIEF-F 的结果特征集 $F''=\{C_{22},\varphi(C_{12}),\text{span},\lambda_2,H,\alpha\}$ 包含的特征个数相等。然而，比较图 6(c)，图 6(d)和图 6(e)可以看出，利用 RELIEF-F 与原始特征集的结果相当，而本文方法的分类结果则明显优于利用原始特征集和 RELIEF-F 的结果。图 6(e)中，4 类地物得到了有效区分，同一类别地物内部误分像素少，地物边缘清晰，图像细节保留完整，分类图与总功率图吻合较好。

6 结束语

在利用 SVM 进行极化 SAR 图像分类的过程中，特征选择能够提高自动化程度，有利于挖掘 SVM 分类潜力和充分利用分类特征所含信息。为此，本文提出一种新的利用 SVM 的 NSVBS 特征选择算法，并将其用于分类的预处理，构成基于 SVM 的极化 SAR 图像分类方法。该分类方法能有效降低 SVM 对自身参数的敏感性。与利用原始特征集和 RELIEF-F 特征选择的分类方法相比，该方法能以更少(或相当)的特征个数，在更广泛的 SVM 参数范围内获得更高的分类精度。对于极化 SAR 图像分类，为获得更好的分类性能，其根本在于充分挖掘回波数据的极化散射信息，这将是本文下一步的研究方向。

致谢 感谢电子科技集团第 38 所张长耀主任提供国内双极化 SAR 数据。

参 考 文 献

[1] Fukuda S and Hirosawa H. Support vector machine classification of land cover: application to polarimetric SAR data. IEEE International Geoscience and Remote Sensing Symposium, Sydney, Australia, July 2001: 187–189.

[2] Cloude S R and Pottier E. An entropy based classification scheme for land applications of polarimetric SAR. *IEEE*

Trans. on Geoscience and Remote Sensing, 1997, 35(1): 68–78.

[3] Vapnik V N. 许建华, 张学工译. 统计学习理论. 北京: 电子工业出版社, 2004: 364–384.

[4] Fukuda S, Katagiri R, and Hirosawa H. Unsupervised approach for polarimetric SAR image classification using support vector machines. IEEE International Geoscience and Remote Sensing Symposium, Toronto, Canada, June 2002: 2599–2601.

[5] Liu Huan and Yu Lei. Toward integrating feature selection algorithms for classification and clustering. *IEEE Trans. on Knowledge and Data Engineering*, 2005, 17(4): 491–502.

[6] Pudil P, Novovicova J, and Kittler J. Floating search methods in feature selection. *Pattern Recognition Letters*, 1994, 15(11): 1119–1125.

[7] Narendra P M and Fukunaga K. A branch and bound algorithm for feature subset selection. *IEEE Trans. on Computers*, 1977, 26(9): 917–922.

[8] Kononenko I. Estimation attributes: analysis and extensions of RELIEF. Proc. 7th European Conference on Machine Learning, Sicily, Italy, April 1994: 171–182.

吴永辉：　男，1976 年生，博士生，研究方向为极化 SAR 信息处理.

计科峰：　男，1974 年生，副教授，主要研究方向为遥感信息处理、SAR 图像理解.

李　禹：　男，1975 年生，博士生，研究方向为遥感信息处理.

郁文贤：　男，1964 年生，教授，博士生导师，主要研究方向为电子系统信息处理与集成.