

基于简约集支持向量机的高光谱影像分类

余旭初¹ 杨国鹏^{1,2} 冯伍法¹ 周 欣³

(信息工程大学测绘学院 郑州 450052)¹ (空军装备研究院情报所 北京 100850)²

(信息工程大学信息工程学院 郑州 450002)³

摘 要 针对高光谱影像支持向量机分类的预测过程中需要花费大量计算时间的问题,提出了一种利用简约集算法提高高光谱影像分类预测效率的方法。采用径向基核函数,使用一对一构造多类支持向量机分类器,通过交叉验证网格搜索法对参数进行模型参数选择,并利用简约集算法来减少分类预测过程计算量。通过高光谱影像分类试验表明,保持支持向量机的泛化能力并不需要使用所有计算得到的支持向量,简约集算法能够在保持分类预测精度的同时,大大提高高光谱影像分类预测的速度。

关键词 高光谱影像,分类,支持向量机,简约集

中图法分类号 TP751 文献标识码 A

Reduced Set Based Support Vector Machine for Hyperspectral Imagery Classification

YU Xu-chu¹ YANG Guo-peng^{1,2} FENG Wu-fa¹ ZHOU Xin³

(Institute of Surveying and Mapping, Information Engineering University, Zhengzhou 450052, China)¹

(Intelligence Institute of Airforce's Equipment Academy, Beijing 100850, China)²

(Institute of Information Engineering, Information Engineering University, Zhengzhou 450002, China)³

Abstract Aiming at the problem of more computational time needed in hyperspectral imagery classification procedure based on support vector machine, a reduced set method was brought forward to heighten hyperspectral imagery classification efficiency. The radial basis kernel function was adopted, one-against-one decomposition algorithm was used to construct multi-class Support Vector Machine classifier and cross validation grid search method was applied to select model parameter. The reduced set algorithm was also used to reduce the computational complexity of predication. Through hyperspectral imagery classification experiment it can be concluded that it does not need to use all support vectors to keep generalization ability of Support Vector Machine. The reduced set algorithm can improve hyperspectral imagery classification predicative efficiency highly and keep classification accuracy at the same time.

Keywords Hyperspectral imagery, Classification, Support vector machine, Reduced set

1 引言

高光谱影像以其丰富的光谱信息给地物分类识别带来巨大机遇,也对传统的分类方法提出了挑战^[1]。统计模式识别与神经网络方法基于大数定理,高光谱影像由于包含的地面细节信息极其丰富,往往无法获得足够数量的训练样本点,影响各类别先验知识的估计;高光谱影像特征维数很高,分类时存在维数灾难^[2]。

支持向量机(Support Vector Machine, SVM)在各领域得到成功应用,引起了模式分析方法的第三次变革,通过核映射,能够使经典的线性方法处理非线性问题^[3]。线性 SVM 以结构风险最小化为归纳原则,在高维特征空间中构造具有低 VC 维的最优分类超平面作为判决面。对于特征空间非线性可分的情况^[4],则通过核函数映射把数据非线性映射到高

维特征空间,在高维特征空间中构造最优分类超平面。SVM 能够有效避免过学习现象,具有良好的泛化能力。

目前,许多学者对基于 SVM 的高光谱影像分类进行了研究。Gualtieri^[5]在 1998 年首次将 SVM 用于高光谱影像分类实验后指出,这种方法能有效地避免 Hughes 现象。夏建涛^[6]对 SVM 的高光谱影像分类性能做了详细实验,指出核函数参数和样本数量对 SVM 的分类性能影响较大。由于 SVM 的决策函数是基函数的线性组合,而且 SVM 的基函数个数基本上随训练样本集规模成线性增长,模型稀疏性有限。这就导致具有海量数据的高光谱影像分类预测过程比较缓慢。Burges 和 Scholkopf 等人提出了简约集算法^[7],使 SVM 的分类器在测试阶段的速度提高了数倍,同时保证分类器的精度仅有微小损失。

本文采用径向基核函数,使用一对一构造多类支持向量

到稿日期:2009-12-15 返修日期:2010-02-28

余旭初(1963—),男,教授,博士生导师,主要研究方向为遥感影像军事应用、图像处理与模式识别;杨国鹏(1982—),男,博士,主要研究方向为图像处理与模式识别、高光谱遥感应用,E-mail: yangguopeng@hotmail.com(通信作者);冯伍法(1963—),男,教授,硕士生导师,主要研究方向为遥感图像判读遥感影像军事应用、图像处理与模式识别;周 欣(1983—),女,博士生,主要研究方向为信号处理及模式识别。

机分类器,通过交叉验证网格搜索法对参数进行模型参数选择,并利用简约集算法来降低分类预测过程的计算复杂度,高光谱影像分类试验证明了该方法的优越性。

2 支持向量机

2.1 支持向量机

SVM 基于统计学习理论的结构风险最小化原则,通过最小化经验风险和置信范围来提高模型泛化能力^[3]。对于训练样本集 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, $y \in \{-1, +1\}$, SVM 优化问题用非线性映射表示为^[4]

$$\begin{aligned} \min_{w, \xi} & \left(\frac{1}{2} (w \cdot w) + C \sum_{i=1}^n \xi_i \right) \\ \text{s. t.} & \begin{cases} y_i [(w \cdot j(x_i)) + b] \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases} \quad i=1, 2, \dots, n \end{aligned} \quad (1)$$

式中, C 为模型规则化系数, $\phi: X \rightarrow F$ 为从输入空间 X 到特征空间 F 的映射, w 为基函数权值向量。

根据 Lagrange 乘子法, SVM 优化问题可用核函数表示为

$$\begin{aligned} \max_{\alpha} & \left(-\frac{1}{2} \sum_{j=1}^n \sum_{i=1}^n \alpha_i \alpha_j y_i y_j k(x_i, x_j) + \sum_{i=1}^n \alpha_i \right) \\ \text{s. t.} & \begin{cases} \sum_{i=1}^n \alpha_i y_i = 0 \\ 0 \leq \alpha_i \leq C \end{cases} \quad i=1, 2, \dots, n \end{aligned} \quad (2)$$

式中, α_i 为训练样本 x_i 对应的权值, $k: X \times X \rightarrow \mathbb{R}$ 为满足 Mercer 条件的核函数, 可以表示为两样本在特征空间投影的内积^[3], 即 $k(x_a, x_b) = \langle \phi(x_a), \phi(x_b) \rangle$ 。

SVM 的判别函数可表示为

$$f(x) = \text{sgn} \left\{ \sum_{x_i \in SV} \alpha_i y_i k(x_i, x) + b \right\} \quad (3)$$

式中, $w = \sum_{x_i \in SV} \alpha_i y_i \phi(x_i)$ 。

SVM 优化问题的复杂度与特征维数无关, 但受训练样本集规模 n 的制约。当训练样本点数目很大时, 计算量剧增, 算法收敛速度慢。许多学者对 SVM 训练算法进行了研究, 并提出了序列最小优化算法、序列过松弛算法、最小二乘法 SVM 等^[2]。

根据核函数理论, 核参数的改变实际上是隐含地改变映射函数从而改变样本数据子空间分布的复杂程度。SVM 中核函数参数与规则化系数的优化方法, 有网格搜索法、梯度下降法、模拟退火算法等。

SVM 是二值分类器, 解决多类分类问题时, 通常将多类问题分解为一系列可直接计算的两类问题, 基于这一系列二类问题求解, 得出最终判别结果。具体实现方法有一对余法 (One Against One, OAO)、一对一法 (One Against Rest, OAR)、二叉树法、纠错输出编码法、有向无环图法等。

本文将采用径向基核函数, 使用一对一或一对余构造多类 SVM 分类器, 使用交叉验证网格搜索法对参数进行优化选择, 并在此基础上利用简约集方法来降低分类过程的复杂度。

2.2 简约集方法

简约集 (Reduced set, RS) 方法通过简化由 SVM 训练的核分类器和其它核机器表达形式来提高核方法的计算效率。假设函数 $f: X \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$ 表示为

$$f(x) = \sum_{i=1}^l a_i k(x_i, x) + b \quad (4)$$

式中, $T_X = \{x_1, \dots, x_l\}$ 为输入空间 $\mathbb{R}^n$ 中的向量集合, $a = \langle a_1, \dots, a_l \rangle^T$ 为实系数权值向量, b 为标量。

在特征空间 F 中, $f(x)$ 可以表示线性函数, 即

$$f(x) = \langle \psi, \phi(x) \rangle + b \quad (5)$$

RS 就是要找到一个简单的新函数, 即

$$\tilde{f}(x) = \sum_{i=1}^m \beta_i k(s_i, x) + b = \langle \underbrace{\phi(x) \cdot \sum_{i=1}^m \beta_i \phi(s_i)}_{\tilde{\psi}}, \phi(x) \rangle + b \quad (6)$$

这里式 (6) 比较简短, 既满足 $m < l$, 又能够近似表示式 (4), 权值向量 $\beta = [\beta_1, \dots, \beta_m]^T$, $T_S = \{s_1, \dots, s_m\}$, $s_i \in \mathbb{R}^n$ 决定了简约核表达式。寻找简约集表达式 (6) 可以表示为优化问题

$$\min_{\beta, T_S} \|\tilde{\psi} - \psi\|^2 = \min_{\beta, T_S} \left\| \sum_{i=1}^m \beta_i \phi(s_i) - \sum_{i=1}^l \alpha_i k(x_i, x) \right\|^2 \quad (7)$$

假设核函数为径向基函数 (Radial Basis Function, RBF), 其表达式为

$$k(x_a, x_b) = \exp(-0.5 \|x_a - x_b\|^2 / \sigma^2) \quad (8)$$

此时, 优化问题 (7) 可以通过迭代贪婪算法获取, 它将优化问题 (7) 转化为一系列的 m 个特征空间的原像问题。

2.3 原像问题

假设 $T_X = \{x_1, \dots, x_l\}$ 为输入空间 $\mathbb{R}^n$ 中的向量集, $X \subseteq \mathbb{R}^n$ 。通过非线性映射 $\phi: X \rightarrow F$ 将数据映射到新的特征空间 F 中, 按照核函数实现。设 $\psi \in F$ 通过以下核展开式给出。

$$\psi = \sum_{i=1}^l \alpha_i \phi(x_i) \quad (9)$$

式中, $\alpha = \langle \alpha_1, \dots, \alpha_l \rangle^T$ 为实系数权值向量, 原像问题就是要找到一个输入向量 $x \in X$, 它的特征空间的像 $\phi(x) \in F$ 能够近似展开点 $\psi \in F$ 。换句话说, 原像问题解决从特征空间 F 映射到输入空间 X , 可以通过以下优化问题实现。

$$\min_x \|\phi(x) - \psi\|^2 = k(x, x) - 2 \sum_{i=1}^l \alpha_i k(x, x_i) + \sum_{i=1}^l \sum_{j=1}^l \alpha_i \alpha_j k(x_i, x_j) \quad (10)$$

假设核函数为径向基函数, 原像问题 (10) 可以转化为优化问题 (11), 具体求解过程参考文献^[7, 8]。

$$\max_x \sum_{i=1}^l \alpha_i \exp(-0.5 \|x - x_i\|^2 / \sigma^2) \quad (11)$$

图 1 为 Ripley 在文献^[9]中使用实验数据的 SVM 与 RS-SVM 的分类结果比较, 共两类 2 维数据, 每类 125 个训练样本。SVM 共求解出 94 个支持向量, RS-SVM 的简约集向量为 10 时就可以获取相似的分类决策面。利用 Riply 测试集对 SVM 和 RS-SVM 进行测试时, 分类错误率都为 0.094。

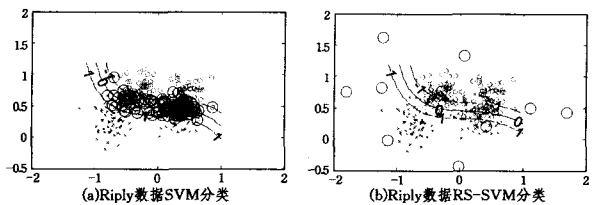


图 1 Riply 数据 SVM 与 RS-SVM 分类试验比较

3 高光谱影像分类实验

实验采用的计算机硬件环境为 Intel Core2 CPU 3.0 GHz, 2.99GHz 内存 3.25GB, 软件环境为 Microsoft Windows XP, MATLAB 7.5。

实验数据为由中科院上海技术物理研究所研制的 PHI 成像光谱仪 1999 年 9 月获取的江苏省常州市影像, 光谱覆盖范围为 0.42~0.85 μm , 共 80 波段, 图像大小为 346 * 512, 数据经过反射率转换。训练样本分布如图 2 所示, 样本数量情况如表 1 所列。



图2 PHI影像训练样本采集情况

表1 该 PHI 数据样本采集情况

类别号	名称	数量
1	房屋	211
2	水体	222
3	土壤	205
4	树木	228
5	农作物	266
6	道路	238

实验前,将各类样本按数量随机等分成两部分作为训练样本和测试样本,训练样本用来进行 SVM 训练,测试样本用来分析 SVM 分类预测精度。实验时,首先将数据各波段特征进行规范化,即使各波段特征值规范化至 $[0,1]$ 内。选用一对一径向基核函数支持向量机,核函数参数和规则化系数通过网格交叉验证获取。

使用 SVM 分类时,重复计算 20 次,OAO-SVM 的训练时间如图 3 所示,平均训练时间为 3.54s,分类预测时间为 154.9s,支持向量数量为 533 个,分类错误率为 0.91%。支持向量数量为 OAO-SVM 中所有二值分类器的支持向量之和。

使用 RS-SVM 分类时,OAO-RS-SVM 分类精度随着简约集向量数的增加迅速提高,当简约集向量数达到一定数量时,分类精度不再提高(如图 4 所示),分类预测的时间随着简约集向量数的增加成线性增加(如图 5 所示),简约集求解的时间也随着简约集向量数的增加而增加(如图 6 所示)。

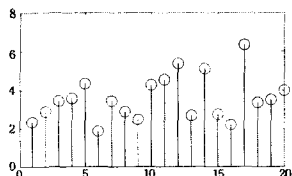


图3 OAO-SVM 的训练时间

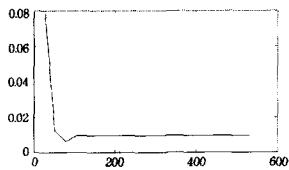


图4 分类错误率随简约集向量数的变化

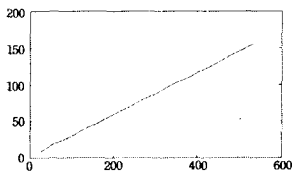


图5 影像分类时间随简约集向量数的变化

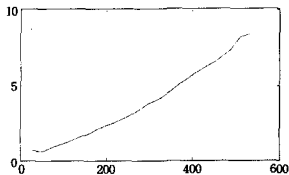


图6 简约集求解时间随简约集向量数的变化

因此,为了提高高光谱影像 SVM 的分类效率,在保持分类精度的同时,通过简约集方法提高 SVM 预测过程中的速

度。当设置简约集为原始支持向量个数的 1/4,即使用简约集向量数为 106 时,粗糙集求解时间为 1.17s,分类预测时间为 31.0s,分类错误率仍为 0.91%,该 PHI 影像分类结果如图 7 所示。



图7 PHI影像 RS-SVM 分类结果

结束语 支持向量机具有良好的泛化能力,在进行高光谱影像分类时具有较高的预测精度,但需要花费大量的时间进行分类预测。由于保持 SVM 泛化能力并不需要使用所有的支持向量,因此使用简约集算法能够有效地减少支持向量数量,从而降低分类预测过程的计算复杂度,提高分类效率。通过高光谱影像分类试验表明,简约集算法能够在保持分类预测精度的同时,大大提高分类预测的速度。

参考文献

- [1] 童庆禧,张兵,郑兰芬. 高光谱遥感——原理、技术与应用[M]. 北京:高等教育出版社,2006
- [2] 杨国鹏. 基于核方法的高光谱影像分类与特征提取[D]. 郑州:信息工程大学,2007
- [3] Shawe-Taylor J, Cristianini N. Kernel Methods for Pattern Analysis [M]. Cambridge University Press,2004
- [4] Vapnik N V. 统计学习理论[M]. 许建华,张学工,译. 北京:电子工业出版社,2004:274-322
- [5] Gualtieri J A, Crompt R F. Support vector machines for hyperspectral remote sensing classification [C] // The 27th AIPR Workshop, Advances in Computer Assisted Recognition. Washington D C, 1998
- [6] 夏建涛. 基于机器学习的高维多光谱数据分类 [D]. 西安:西北工业大学,2002
- [7] Scholkopf B, Knirsch P, Smola C, et al. Fast approximation of support vector kernel expansions and an interpretation of clustering as approximation in feature spaces[C]//DAGM. Springer, 1998:124-132
- [8] Kwok J T, Tsang I W. The pre-image problem in kernel methods[C]// Proceedings of the Twentieth International Conference on Machine Learning. Washington D C, USA, 2003: 408-415
- [9] Ripley B D. Neural networks and related methods for classification [J]. Journal of the Royal Statistical Society, Series B, 1994, 56:409-456