

感兴趣区域遥感图像分类与支持向量机应用研究

曾联明^{1,2,3}, 吴湘滨¹, 刘鹏³

ZENG Lian-ming^{1,2,3}, WU Xiang-bin¹, LIU Peng³

1.中南大学 地学与环境工程学院,长沙 410083

2.佛山科学技术学院 信息中心,广东 佛山 528000

3.解放军理工大学 网络研究中心,南京 210016

1.School of Geosciences and Environmental Engineering, Central South University, Changsha 410083, China

2.Information and Educational Technology Center, Foshan University, Foshan, Guangdong 528000, China

3.MilGrid Research Center, The PLA University of Science and Technology, Nanjing 210016, China

E-mail: zenglm@163.com

ZENG Lian-ming, WU Xiang-bin, LIU Peng. Research on SVM and its application of remote sense image classification for regions of interest. *Computer Engineering and Applications*, 2009, 45(6): 243-245.

Abstract: A classification method and model is proposed based on SVM for remote sense image. By selecting Regions Of Interest (ROI) from the 1:50 000 TM image of Tangshan city area, extract the feature of greenbelt, public lands, building and so on, the parameters of C and γ are achieved by cross validation method, with these textures to train and parameters to classify the RS image, the fact shows that the classification method based on SVM has a high accuracy and a fast, stably efficiency.

Key words: remote sense; image classification; Support Vector Machine(SVM); Regions Of Interest(ROI)

摘要:提出了基于SVM的遥感图像分类方法并构建了分类模型,该方法以唐山1:50 000 TM局部图为分类数据来源,由用户选择感兴趣的区域,分别提取该区域绿地、公共用地和房屋的图像特征,并以此为训练样本进行训练,采取交叉校验的方法获得SVM的最优惩罚因子 C 和间隔 γ 参数进行图像分类。实验结果表明,此分类方法准确率高、稳定快捷,是SVM在遥感图像分类中的一个很好的应用。

关键词: 遥感; 图像分类; 支持向量机; 感兴趣区域

DOI: 10.3778/j.issn.1002-8331.2009.06.070 文章编号: 1002-8331(2009)06-0243-03 文献标识码: A 中图分类号: TP315

1 绪言

遥感图像计算机分类是遥感技术应用的一个重要组成部分^[1]。目前常用的遥感图像分类方法是以贝叶斯(Bayes)的主观概率理论为代表的最大似然法,它的前提条件要求图像数据呈正态分布,而且最主要的问题是反映不出对知识的不知道信息^[2],一次分类不满意,就要重新选择训练样本或者增加分类数重新分类,效率太低,且不能保证精度一定会提高^[3]。遥感图像的兴趣区域是图像的一部分,可以以图元像素或者根据阈值等其它方式来选择。这些区域可以是不规则的形状,也可以是几个不同形状的组合,被广泛地使用于分类的地物特征抽取、统计和其他的操作中。而基于传统统计学的遥感图像分类方法,需要有足够多的样本,但遥感图像分类总是在样本有限的情况下进行的,所以难以取得理想的分类效果^[4]。

支持向量机简称SVM(Support Vector Machines)是Vapnik等人根据统计学理论提出的一种机器学习方法^[5],它集成了最大间隔超平面、Mercer核、凸二次规划和松弛变量等多项技

术。支持向量机由于较好地解决了小样本、非线性、高维数、局部极小点等实际问题,在若干具有挑战性的应用中,获得了目前为止最好的性能^[6]。支持向量机已经逐渐成为解决模式分类问题的首选工具。本文在遥感图像感兴趣的有限区域进行数据采集,并以此作为分类的样本进行实验和分类,是基于SVM的方法在遥感图像分类中的一个应用。

2 支持向量机原理

SVM的机理:寻找一个满足分类要求的最优分类超平面,使得它在保证分类精度的同时,既尽可能多的将两类数据点正确的分开,同时使得该超平面两侧的空白区域最大化,即使分开的两类数据点距离分类面最远。

同一个训练样本可以有被不同超平面分类的情况,当超平面的空白区域Margin最大时,超平面就是最优分类超平面^[7]。设训练样本集:

$$(x_1, y_1), \dots, (x_l, y_l), x_i \in R^n, y_i \in \{-1, +1\} \quad (1)$$

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.40473029)。

作者简介:曾联明(1973-),男,博士研究生,主研方向:网络计算技术、信息集成、地理信息;吴湘滨(1962-),男,教授,博士生导师,主研方向:地质灾害评价、预测与防治、地理信息;刘鹏(1970-),男,教授,博士,IEEE、GGF、CGF会员,硕导,主研方向为高性能计算、网络技术。

收稿日期:2008-01-08 修回日期:2008-03-31

x_i 为训练样本输入, $i=1, \dots, l, y_i$ 为期望输出, 式(2)为分类面方程。

$$(w \cdot x) + b = 0 \quad (2)$$

为使分类面对所有样本正确分类并且具备分类间隔, 使得:

$$\begin{cases} (w \cdot x_i) + b \geq 1 & y_i = 1 \\ (w \cdot x_i) + b \leq -1 & y_i = -1 \end{cases} \Leftrightarrow y_i [(w \cdot x_i) + b] \geq 1 \quad (i=1, 2, \dots, l) \quad (3)$$

可以计算出, 训练数据集到一给定的超平面的分类间隔为:

$$\rho(w, b) = \min_{\{x_i, y_i=1\}} \frac{w \cdot x_i + b}{\|w\|} - \max_{\{x_i, y_i=-1\}} \frac{w \cdot x_i + b}{\|w\|} = \frac{2}{\|w\|} \quad (4)$$

根据 SVM 对优化超平面的定义, 可以看出对该平面的求解问题可以简化为: 在满足条件式(3)的情况下, 计算能最大化 $\rho(w, b)$ 的超平面的法向量 w 和偏移量 b , 即求分类间隔 $\frac{2}{\|w\|}$ 的最大值, 使间隔最大等价于使 $\|w\|^2$ 最小。使式(3)等号成立的样本叫做支持向量, 满足条件(3)且使 $\frac{1}{2} \|w\|^2$ 最小的分类面就叫做最优分类面。

Vapnik 等人证明: 分割超平面的法向量 w_0 是所有训练集向量的线性组合。即 w_0 可以描述为:

$$w_0 = \sum_{i=1}^l (a_i^0 y_i) x_i \quad (a_i^0 \geq 0), i=1, \dots, l \quad (5)$$

在多数情况下式(5) w_0 的展开式中, 系数 a_i^0 为零值, 而非零值的 a_i^0 对应的 x_i 就称为支持向量 SV。引入拉格朗日乘子 $\alpha_i, i=1, 2, \dots, l$, 并定义

$$w(\alpha) = \sum_{i=1}^l \alpha_i y_i x_i \quad (6)$$

使用对偶定理把上述问题转化为其对偶问题:

$$\text{Max} W(\alpha) = \sum_i \alpha_i - \frac{1}{2} w(\alpha) \cdot w(\alpha) \quad (7)$$

其中 $\alpha_i \geq 0, \sum_i \alpha_i y_i = 0$ 。

解上述问题后得到的最优决策函数是:

$$f(x) = \text{sgn} \left(\sum_{i=1}^n a_i^* y_i (x_i \cdot x) + b^* \right) \quad (8)$$

在线性不可分的情况下, 可以在条件(3)中增加一个松弛项 $\xi_i \geq 0$, 成为:

$$y_i [(w \cdot x_i) + b] - 1 - \xi_i \geq 0, i=1, 2, \dots, n \quad (9)$$

将目标改为求:

$$(w, \xi) = \frac{1}{2} \|w\|^2 + C \left(\sum_{i=1}^n \xi_i \right) \quad (10)$$

的最小且 $0 \leq \alpha_i \leq C$, 其中, C 是惩罚因子, 即综合考虑最少错分样本和最大分类间隔, 这样就得到广义最优分类面。

对于非线性问题, 只需要将输入向量非线性映射到一个更高维的特征空间, 然后再构造最优分类超平面。不必知道具体的映射函数 $\phi(x_i)$ 的表达式, 因为在这个高维空间中只涉及到内积运算, 若 $K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$, 则称 $K(x_i, x_j)$ 为内核函数, 一个函数是核函数的条件由 Mercer 定理给出。而相应的最优决策函数变为:

$$f(x) = \text{sgn} \left(\sum_{i=1}^n a_i^* y_i K(x_i \cdot x) + b^* \right) \quad (11)$$

常用的核函数有: 多项式核函数, 径向基核函数, Sigmoid

核函数。

由于高斯核函数是局部性非常强的函数, 可逼近任意非线性函数^[8], 本文采用基于二叉树的多类 SVM 分类方法和径向基核函数进行研究, 径向基核函数表示为:

$$K(x \cdot x_i) = \exp \left(-\frac{1}{2\sigma^2} \|x - x_i\|^2 \right)$$

3 基于 SVM 的遥感图像分类方法

3.1 图像分类器模型

常用的 SVM 多类分类方法有一对一(1-a-1)和一对多(1-a-r)两种, 遥感图像分类器是基于二叉树的多类 SVM 分类器提出来的。

基于二叉树的多类 SVM 对于 K 类的训练样本, 训练 $K-1$ 个支持向量机^[9]。第一个支持向量机以第一个样本为正样本, 将第 2, 3, ..., K 类训练样本作为负的训练样本训练 SVM1; 第 i 个支持向量机以第 i 个类样本为正的训练样本, 将第 $i+1, i+2, \dots, K$ 类训练样本作为负的训练样本训练 SVM i , 直到 $K-1$ 个支持向量机将以第 $K-1$ 类样本作为正样本, 第 K 类样本为负样本训练 SVM($K-1$)。图 1 为基于 SVM 的二叉树多类遥感图像分类器结构。

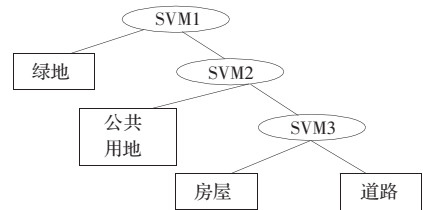


图 1 基于 SVM 的二叉树多类遥感图像分类器结构

二叉树方法可以避免传统方法的不可分情况, 并只需构造 $k-1$ 个 SVM 分类器, 测试时并不一定需要计算所有的分类器判别函数, 从而可节省测试时间。在实际的遥感图像分类过程中, 在分类类别比较少并且不强调时间的情况下也可以采取只训练一个支持向量机的方法分别来进行分类, 即将需要分类的样本都作为正样本, 其它都为负样本, 分别进行。

3.2 基于 SVM 分类方法的图像分类流程

基于二叉树的遥感图像感兴趣区域多类 SVM 分类过程主要包括: 图像预处理、感兴趣区域属性特征及训练样本的提取、数据标准化及归一化处理、 C, γ 参数设置、基于 SVM 图像分类和分类精度评价几个部分。

训练样本选取原则要充分考虑各种地物的光谱、结构和纹理特征, 因地制宜地进行选择^[10]。感兴趣区域的选择方法可以满足训练样本的选择要求, 可以是单个的多边形所包含的区域, 也可以是多边形、点、矢量等的组合区域。遥感图像的感兴趣区域样本选择在 ENVI4.1 软件中进行, ENVI4.1 支持多个感兴趣区域的选择, 选择的区域可以根据用户需要进行定义, 并且在任何一个主图像窗口、滚动窗口以及缩放窗口中进行绘制。为了适合一个特殊的图元像素的阈值, 感兴趣区域可以扩展到该图元像素的临近像素, 保证了所选择的样本区域特征信息的广泛性, 其流程如图 2 所示。

预处理阶段, 是实施图像分类的关键阶段, 感兴趣样本选择后必须转换成可以被分类识别器所能识别的数据格式, 所以必须编写专门的数据格式转换程序来实现。数据的归一化处理

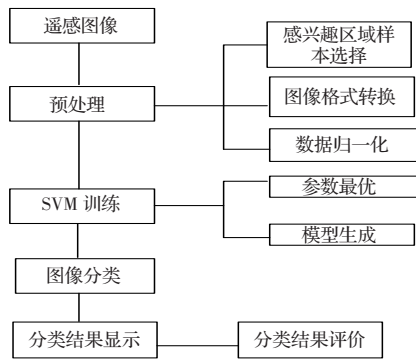


图2 SVM 感兴趣区域图像分类流程

在 SVM 执行分类处理中非常的重要^[1],一方面,可以避免小数值范围的属性的值被更多大数值范围的属性的数值所支配;另一方面可以避免统计过程中遇到的计算困难。在实际数据处理过程中通常建议将属性的值归一到 $[-1, 1]$ 或 $[0, 1]$ 的范围。SVM 样本训练中参数的优化主要是 (C, γ) 参数的优化,也是一个决定分类效果的重要因素,在图像分类中采用的基于“网格搜索”的交叉验证方法为 (C, γ) 优化参数的获取,提供了途径。

4 实验结果分析

图3为原始图像数据,其分辨率为1:50 000,图像大小为 175×572 。图4使用SVM分类出来的绿地(原图白色部分),图5使用SVM分类出来的公共用地(原图绿色部分),图6使用SVM分类出来的房屋建筑(原图粉红色部分)。实验硬件平台为P4-MCPU,主频1.8 GHz,内存512 MB,在.net 环境下以libsvm-2.84为基础编写了图像分类器程序,从图中可以看到,图像的分类效果非常理想。

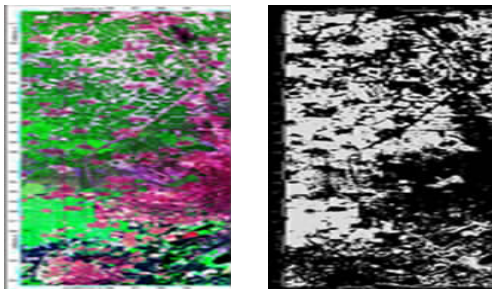


图3 唐山部分地区 (1:50 000TM 图局部) 图4 使用 SVM 分类出来的绿地 (原图绿色部分)

5 结束语

SVM 是一种基于统计学习理论的模式识别方法,SVM 以

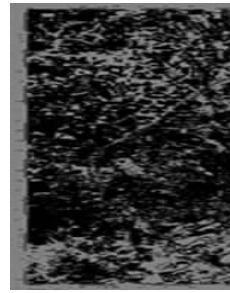


图5 使用 SVM 分类出来的公共用地 (原图白色部分)

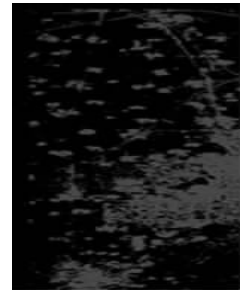


图6 使用 SVM 分类出来的房屋建筑 (原图粉红色部分)

有限的样本信息,根据结构风险最小化原则,在模型的复杂性和学习能力之间寻求最佳折中,以期达到最好的泛化能力。它将优化问题转化为求解一个凸二次规划的问题,二次规划所得的解是唯一的且为全局最优解,这样就不存在一般神经网络的局部极值问题。

参考文献:

- [1] 孙家柄.遥感原理与方法[M].武汉:武汉大学出版社,2003.
- [2] Richards A.Remote sensing digital image analysis:an introduction[M]. 2nd ed.Berlin:Springer-Verlag,1993.
- [3] 邓文胜,邵晓莉,刘海等.基于证据理论的遥感图像分类方法探讨[J].遥感学报,2007,11(4):568-573.
- [4] 惠文华.基于支持向量机的遥感图像分类方法[J].地球科学与环境学报,2006,8(2).
- [5] Vapnik V N.The nature of statistical learning theory[M].New York: Springer-Verlag,1995.
- [6] 张学工.关于统计学习理论与支持向量机[J].自动化学报,2000,26(1):32-42.
- [7] Cristianini N,Shawe-Taylor J.An introduction to support vector machines and other Kernel-based learning methods[M].[S.l.]:House of Electronics Industry,2005.
- [8] Kamran U K,Yang Jian.Polarimetric synthetic aperture radar image classification by a hybrid method[J].Tsinghua Science and Technology,2007,12(1):97-104.
- [9] 孟媛媛.一种新的基于二叉树的 SVM 多类分类方法[J].计算机应用,2005,25(11):2653-2657.
- [10] 张锦水,何春阳,潘耀忠,等.基于 SVM 的多源信息复合的高空间分辨率遥感数据分类研究[J].遥感学报,2006,10(1):49-57.
- [11] Hsu C W,Chang C C,Lin C J.A practical guide to support vector classification[EB/OL].http://www.csie.ntu.edu.tw/~cjlin/.

(上接 242 页)

参考文献:

- [1] Qiao F,Wang X,Yu L.Optimizing aggregation level for ITS data based on wavelet decomposition[C]//82nd Transportation Research Board Annual Meeting CD-ROM,Washington D C,January 2003.
- [2] 于雷,乔凤祥,徐力,等.基于抽样技术的 ITS 数据存档策略[J].交通运输系统工程与信息,2003,3(4):16-26.

- [3] 肖扬,鲁凌云,高爽,等.基于二维离散小波变换的智能交通系统数据去噪声压缩[J].北京交通大学学报:自然科学版,2004,28(5):1-5.
- [4] 范莹,须德,郝雪峰.C/S 和 B/S 模式相结合的 ITS 数据再获得系统[J].计算机工程与应用,2004,40(35):208-210.
- [5] 周三多.管理学[M].北京:高等教育出版社,2000:268-269.
- [6] 耿彦斌,于雷,赵慧.ITS 数据质量控制技术及应用研究[J].中国安全科学学报,2005,15(1):82-87.