

基于支持向量机的高维多光谱数据分类^{*})

High Dimensional Multispectral Data Classification by Support Vector Machine

何明一 夏建涛

(西北工业大学电子工程系 西安710072)

Abstract Support vector machine (SVM) is recently developed as a novel learning machine based on statistic learning theory. It is well generalized even when the training samples set is small. The difficulties in high dimensional multispectral data classification are analyzed, which are small training samples set, high dimension and serious Hughes phenomenon etc. So it is a good idea to apply support vector learning theory in high dimensional multispectral data classification. Support vector classifier (SVC) is constructed in this paper. The classification accuracy of SVC and traditional algorithms used in multispectral data classification are compared with computer generated data and multispectral data. Experiments results show that the SVC can overcome the difficulties met in high dimensional multispectral data classification easily. It has higher classification accuracy than traditional algorithms, especially when the training samples set is small.

Keywords Support vector machine, Neural networks, High dimensional multispectral image processing, Pattern classification

1 引言

多光谱技术是人们获取、感知地球表面信息的一种重要手段^[1]。随着传感器技术的进步,新型多光谱传感器光谱分辨率达到了纳米级,具有几百个(100~300)光谱波段,这为人们提供了丰富的地物信息。然而由于其数据维数高、学习样本数量常常有限,使得高维多光谱数据分类面临许多新的课题。传统模式识别方法在处理高维多光谱数据时,面临着巨大的困难:(1)基于经典统计学的模式识别方法只有当训练样本数量趋于无穷时,性能才能达到最优。当样本数量较少时,通常难以取得很好的结果;(2)传统方法在高维数据分类时 Hughes 现象十分严重^[2];(3)基于 Bayes 理论的方法需要对数据的概率分布做出假设,且当假设与数据真实概率分布不相符合时,分类精度急剧恶化;(4)神经网络分类器已经被应用于多光谱数据分类^[3,4],取得了比 Bayes 方法更好的结果,但神经网络存在训练速度慢,易陷入局部极小点等缺陷。

70年代,Vapnik 等建立了统计学习理论的基本体系,系统地研究了有限样本情况下的统计学习问题。近几年,在这一理论框架下产生出了“支持向量机(Support Vector Machine-SVM)”机器学习方法^[5]。SVM 以其完备的理论体系和强的推广能力,受到越来越多的重视^[7,8]。SVM 已经在模式识别、回归估计、密度估计等领域取得了很好的结果^[6-8]。

本文把 SVM 应用于高维多光谱数据分类,构造了支持向量分类器(SVC),克服了传统算法在学

习样本少、数据维数高时表现出来的推广性差、Hughes 现象严重和易过学习等缺陷,提高了分类精度,并采用人工数据和多光谱数据,通过与四种传统分类方法的对比,研究了 SVC 的分类特性。实验结果表明,SVC 能够适应高维数据分类,分类精度明显优于传统分类器,尤其当学习样本较少时,SVC 的优势更加明显。

2 支持向量分类器(SVC)

给定一个内积空间 Z , 一个超平面函数集 $\{z \in Z; \langle W, z \rangle + b = 0\}$ ($\langle W, z \rangle$ 表示两个向量的内积运算)和一组独立同分布的观测数据 $\{x_i, y_i\} (i=1, 2, \dots, l)$, 其中 $x_i \in R^N, y_i \in \{-1, +1\}$ 。SVC 构造具有以下形式的判决函数:

$$f(x) = \text{sgn} \left[\sum_{i=1}^l a_i y_i K(x_i, x) + b \right] \quad (1)$$

其中 $a_i \geq 0, b$ 为实数, $K(\cdot, \cdot)$ 是符合 Mercer 条件^[6]的核函数。

SVC 的核心思想就是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具有低 VC 维的最优分类超平面^[9]。通过综合考虑经验风险和推广能力,寻求使分类风险上界最小的函数作为判决函数。其过程可简要描述如下:

$$\min_{w, \xi} r(W, \xi) = \frac{1}{2} \langle W, W \rangle + C \sum_{i=1}^l \xi_i \quad (2)$$

$$\text{s. t.} \begin{cases} y_i [\langle W, \Phi(x_i) \rangle + b] \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases} \quad i=1, 2, \dots, l \quad (3)$$

其中 $\Phi(\cdot)$ 是非线性映射函数, ξ_i 是松弛因子。

^{*}) 本研究得到国防973计划和教育部博士点基金资助。何明一 教授, 博士生导师, 主要研究领域: 神经网络与智能信息处理, 图像工程与可视化, 以及三维信息获取与处理等。夏建涛 博士生, 主要研究领域: 神经网络与遥感成像。

公式(2)的对偶问题是如下的 QP 优化问题,

$$\begin{aligned} \max_{\alpha} Q(\alpha) = & -\frac{1}{2} \sum_{j=1}^l \sum_{i=1}^l \alpha_i \alpha_j y_i y_j K(x_i, x_j) + \sum_{i=1}^l \alpha_i \quad (4) \\ \text{s. t. } & \begin{cases} \sum_{i=1}^l y_i \alpha_i = 0 \\ 0 \leq \alpha_i \leq C \\ \alpha_i [\langle W, \Phi(x_i) \rangle + b + \xi_i - 1] = 0 \quad (k, k, t) \end{cases} \quad i=1, 2, \dots, l \end{aligned} \quad (5)$$

其中核函数 $K(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle$ 。

由方程(5)中的 (k, k, t) 条件^[9]可以看出,只有与判决边界最近或被错误分类的样本对应的 $\alpha_i > 0$, 其它样本对应的 $\alpha_i = 0$ 。我们把 $\alpha_i > 0$ 的样本称为支持向量(Support Vector—SV)。超平面的权向量 W 可表示为:

$$W = \sum_{x_i \in \text{SV}} \alpha_i y_i \Phi(x_i) \quad (6)$$

其中 SV 表示支持向量集合。

$$\text{判决函数为: } f(x) = \text{sgn} \left[\sum_{x_i \in \text{SV}} \alpha_i y_i K(x_i, x) + b \right] \quad (7)$$

3 支持向量分类器的性能

我们分别用人工数据、64波段多光谱数据和220波段高维光谱数据,把 SVC 与4种传统算法(最小距离分类器^[1]—MD,最大似然分类器^[1]—ML,多层前向神经网络分类器^[5]—NNC,径向基神经网络分类器^[6]—RBFNNC)进行对比研究。

3.1 人工数据分类

I : 高斯数据分类

产生两类高斯分布的数据,其均值和方差矩阵分别为:

$$\begin{aligned} M_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}, \\ M_2 = \begin{bmatrix} 3 \\ 1 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix} \end{aligned}$$

为了对比各分类器对训练样本数量的适应能力,我们设定两种情况:① 训练样本数量为200;②

训练样本数量为20。两种情况下测试样本均为2000。实验重复5次,求平均精度。各个分类器的测试分类精度如表1所示。

表1 各分类器对高斯分布数据分类精度的对比

分 类 器	MD	ML	NNC	RBFNNC	SVC
训练样本 200	83.30	88.83	88.82	88.84	89.82
训练样本 20	72.90	74.00	70.79	72.49	85.08

II : 双螺旋数据分类

我们构造了双螺旋数据,如图1所示。实验中取训练样本100个,测试样本300个。



图1 双螺旋数据

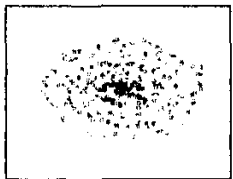


图2 加10%高斯噪声的双螺旋数据

为了检测各分类器对噪声的适应能力,实验数据分为两种情况:① 不加噪声的测试样本(如图1所示);② 加10%高斯噪声的测试样本(如图2所示)。各分类器形成的判决面分别如图3(a~e)所示。分类器的测试分类精度在表2中给出。

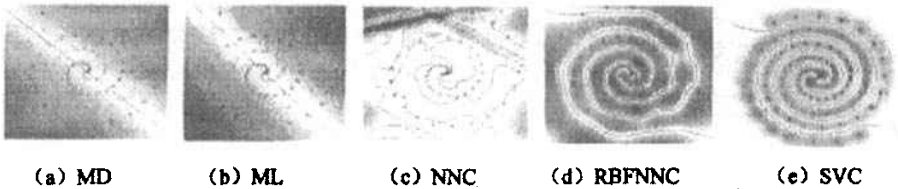


图3 分类器对双螺旋数据分类时形成的分类面

表2 各分类器对双螺旋数据的分类精度对比

分 类 器	MD	ML	NNC	RBFNNC	SVC
训练样本 无噪声	57.67	56.00	89.67	98.33	99.67
测试样本 加噪声	56.33	55.67	77.99	91.00	94.00

3.2 多光谱数据实验

I : 64波段多光谱数据分类

下面的实验采用国家863-308提供的64波段多光谱数据。为了研究各分类器对训练样本数量的适应能力,我们分别选训练样本较多和较少两种情况

进行对比.表3给出了本次实验选取的地物类型和样本数量.图4是这两种地物的平均光谱响应曲线.

表3 样本描述

地物类别	数据1		数据2	
	训练样本	测试样本	训练样本	测试样本
Grass	158	946	70	946
Unknown Vegetable	241	1503	70	1503

各分类器对数据1和数据2进行分类的精度如表4.

表4 各分类器对64波段多光谱数据分类的精度对比

分类器		MD	ML	NNC	RBFNNC	SVC
分类精度(%)	数据1	89.34	91.67	94.20	94.61	94.80
	数据2	89.38	62.72	84.36	91.67	93.54

I :220波段多光谱数据分类

下面实验的数据是1992年用 AVIRIS 多光谱扫描仪获得的220波段高维多光谱数据.我们选取两种地物类型的数据,地物类别和样本数量如表5所示.各分类器对数据1和数据2进行分类的精度在表6中给出.与实验1类似,我们分别选训练样本较多和较少两种情况进行对比,图5是这两种地物的平均光谱响应曲线.

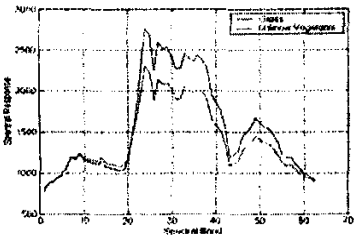


图4 两种地物的平均光谱响应

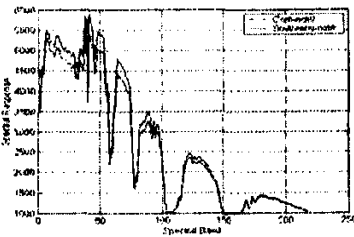


图5 表5中两种地物类别的平均光谱响应

3.3 结果分析

根据人工数据和多光谱数据实验结果,我们有以下结论:

(1)无论学习样本数量的多少,数据的概率分布如何,SVC 都取得了比传统分类器更好的结果.当

学习样本数目较少时,SVC 的优势更加明显(表1,表4,表6).这充分说明 SVC 在学习样本数目较小的情况下,具有很好的推广能力.

表5 样本描述

地物类别	数据1		数据2	
	训练样本	测试样本	训练样本	测试样本
Corn-notill	300	485	100	685
Soybeans-notill	300	445	100	645

表6 各分类器对220波段高维多光谱数据分类的精度对比

分类器		MD	ML	NNC	RBFNNC	SVC
分类精度(%)	数据1	74.73	83.01	98.49	98.06	99.35
	数据2	73.90	80.83	89.98	94.21	96.51

(2)对比表1和表2,当数据不服从高斯分布时,MD 和 ML 分类器的性能大幅度降低.这是因为它们分类时假设数据服从高斯分布.NNC、RBFNNC 和 SVC 都不受数据概率分布的影响.

(3)MD(一阶分类器)与 ML(二阶分类器)不能在数据空间形成复杂的非线性分类面(图3(a,b)),所以对具有复杂空间分布的数据分类的精度不高(表2,4,6).NNC、RBFNNC 和 SVC 都可以在数据空间形成高度复杂的分类面(图3(c-e)),所以能够适应复杂空间分布数据的分类.

(4)表2结果显示,当测试样本含有噪声时,NNC 和 RBFNNC 的分类精度有较大幅度的降低,而 SVC 的分类精度降低幅度较小,这说明 SVC 具有强的推广能力.

(5)NNC 对训练样本数量需求很大.当训练样本较少时,其分类精度有较大幅度的降低(表1,2,4,6).

总结与展望 针对高维多光谱数据分类面临的困难(训练样本少、数据维数高、Hughes 现象严重),本文把统计学习理论和支持向量机用于多光谱数据分类,构造了支持向量分类器(SVC),有效地克服了高维数据分类中的难题.采用人工数据、64波段多光谱数据和220波段高维光谱数据对比研究了 SVC 和最常用的4种分类器(MD,ML,NNC,RBFNNC)的分类性能.实验结果表明,SVC 比传统分类器具有更好的分类精度,尤其是当训练样本数目较少时,SVC 的优势更加突出.这充分说明基于小样本统计学的 SVC 在训练样本少时,仍具有很好的推广能力,非常适合于高维多光谱数据分类.

参考文献

1 Richards J A, Jia Xiuping. Remote sensing digital image analysis. Springer-Verlag, Berlin Heidelberg. 1999 (下转第51页)

(i', j') 的值就等于 $f(m, n)$ 的红色分量的 Newton-Thiele 型插值曲面的第 (i, j) 块 $I_{i,j}$ 在 $(\Delta x, \Delta y)$ 坐标处的值。

利用同样的方法, 可以求出绿色分量 $G(i, j)$ 和蓝色分量 $B(i, j)$ 的值。

4. 实验结果

我们在 Pentium 655 上用 Dephi 5.0 实现了该算法, 并验证了该方法的可行性。以 256×256 的图



图1 Panda 原图



图2 利用本文放大(307×460)

Panda. bmp 为例, 分别对其进行放大和缩小。如图1—图3所示。图2的执行时间为535ms。图3的执行时间为172ms。同文[2]和文[4]相比, 速度明显提高。



图3 利用本文缩小(205×128)

结论 从实验结果可以看出, 利用 Newton-Thiele 型插值曲面进行缩放效果是好的, 是一种既快速又能保证质量的图像缩放方法, 这对于在多媒体和互联网上图像的处理有一定的意义。

参考文献

- 1 朱志刚, 林学闯, 石定机译. 数字图像处理. 北京: 电子工业出版社, 1998. 97~117
- 2 杨朝霞, 等. 用 B 样条的尺度关系来实现图像任意精度的放大缩小[J]. 计算机辅助设计与图形学学报, 2001, 13(9): 824~827
- 3 Durand C X, Faguy D. Rational zoom of bitmaps using B-spline interpolation in computerized 2-D animation. Computer Graphics Forum. 1990, 9(1): 27~37
- 4 孙庆杰, 等. 一种基于 Bezier 曲面的图像放大方法[J]. 软件学报, 1999, 10(6): 570~574
- 5 檀结庆, 胡敏, 刘小平. 有理曲面的三维重建[J]. 计算机应用 2000, 20(Suppl.): 57~59
- 6 Tan Jieqing, Fang Yi. Newton-Thiele' rational interpolants[J]. Numerical Algorithms, 2000, 24: 141~157
- 7 Tan J. Bivariate blending rational interpolants[J]. Approx. Theory &its Application, 1999, 15(2): 74~83
- 8 Siemaszko W. Thiele-type branched continued fractions for two variable functions[J]. Comput. Appl. Math., 1983, 9: 137~153

(上接第96页)

- 2 Landgrebe D. Some fundamentals and methods for hyperspectral image data analysis. SPIE International Symposium on Biomedical Optics (Photonics West), San Jose California, Jan. 1999. 23~29
- 3 Heermann P D, Khazenie N. Classification of multispectral remote sensing data using a back-propagation neural network. IEEE Trans. on Geosci and Remote Sensing, 1992, 30(1): 81~88
- 4 He Mingyi, Bogner R E, et al. Classification of multispectral imagery by using neural network with binary data. Australia First

- Conference on Artificial Neural Networks, Sydney, Jan. 1990
- 5 Haykin S. Neural networks: A comprehensive foundation. Prentice Hall, 2001
- 6 Vapnik 著, 张学工译. 统计学习理论的本质. 清华大学出版社, 2000
- 7 Schölkopf B. Support vector learning: [PhD Thesis]. Berlin, 1997
- 8 Burges CJC. A tutorial on support vector machines for pattern recognition. Knowledge Discovery and Data Mining, 1998, 2(2)
- 9 陈宝林. 最优化理论与算法. 清华大学出版社, 1996