

基于支持向量机的恒星光谱分类研究^{*})

邢 飞 郭 平

(北京师范大学计算机科学系 北京 100875)

摘 要 本文提出了一种新的恒星光谱分类技术。由于恒星光谱数据通常含有较高的噪声,首先我们采用小波变换的方法对原始数据进行降噪,然后采用支持向量机技术完成对光谱数据的最终分类。实验结果显示,使用这种组合分类方法的分类效果要优于使用支持向量机结合主分量分析降维技术的分类方法。

关键词 小波降噪,主分量分析,恒星光谱数据,支持向量机

1 引言

恒星光谱分类是天体光谱自动识别中的重要组成部分,是恒星和星系天文学中的一个重要工具。由于光谱数据数量十分巨大,手工进行光谱分类是不现实的。因此,为了达到光谱数据的自动处理,急需研究快速自动的恒星光谱分类技术。

近年来,小波变换由于其出色特性已成为众多学科关注的焦点,它具有良好的时频局域化特性,能通过伸缩和平移对信号进行多分辨率分析,能聚焦到对象的任意细节,因而是一种很好的信号分析手段。

在本文中我们首先采用小波变换的方法对原始数据进行降噪,然后将降噪后的数据作为支持向量机的输入完成最终的分类。支持向量机(Support Vector Machines,简称 SVM)是 Vapnik 等人^[1,2]提出的一类新型机器学习方法,由于其出色的学习性能,该技术已成为机器学习界的研究热点,并在很多领域都得到了成功的应用。由于支持向量机在高维小训练样本情况下有着很好的处理能力,因此这个特性可以很好地应用于恒星光谱分类中。

2 基本原理

2.1 小波变换原理

小波变换可同时在时域和频域上分析信号的局部特性。平方可积函数

$f(t) \in L^2(R)$ 的连续小波变换定义为

$$\begin{aligned} WT_f(a, b) &= \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f(t) \Psi\left(\frac{t-b}{a}\right) dt \\ &= \langle f(t), \Psi_{a,b}(t) \rangle \end{aligned} \quad (1)$$

其中,小波变换的核函数 $\Psi_{a,b} = \frac{1}{\sqrt{a}} \Psi\left(\frac{t-b}{a}\right)$ 是母小波 $\Psi(t)$ 的时间平移 b 和尺度伸缩 a 的结果。

应用小波变换降噪主要是利用小波变换能较好

地表示信号的局部结构特征,以及信号局部结构特征下所表现的奇异性不同于噪声所表现的奇异性。在本文实验中我们直接利用 Mallat 算法将信号分解为高频和低频信息,设定一个截断尺度,使频率高于此尺度下的小波空间向量全部置为零,然后进行信号重构即可达到去噪的作用。在小波系数的取舍问题上,采用 Donoho 提出的通用阈值算法^[3]。通用阈值 T 由下式定义:

$$T = \sqrt{2 \log_2 S} \quad (2)$$

其中对于小波包变换计算, $S = N \log_2 N$; 对于离散小波变换计算, $S = N$ 。Donoho 还提出了选择系数的软硬阈值法。在硬阈值法中,所有变换系数的绝对值与通用阈值 T 作比较。如果一个系数值小于通用阈值 T ,则等于 0,

$$\text{硬阈值 } a_j = \begin{cases} 0 & |a_j| < T \\ a_j & |a_j| \geq T \end{cases} \quad (3)$$

上式中, a_j 代表小波域上的系数。软阈值法的系数由下式确定:

$$\text{软阈值 } a_j = \begin{cases} 0 & |a_j| < T \\ |a_j| - T & |a_j| \geq T \end{cases} \quad (4)$$

2.2 支持向量机原理

SVM 是以结构化风险最小化(SRM)代替常用的经验风险最小化(ERM)作为优化准则,可以在理论上取得更好的推广性能。为了最小化期望风险的上界,SVM 在固定学习机经验风险的条件下最小化 VC 置信度。SVM 的基本思想是对于非线性可分样本,将其输入向量经非线性变换映射到另一个高维空间 Z 中,在变换后的空间中寻找一个最优的分界面(超平面),使其推广能力最好。以两类模式的分类为例说明其基本原理。

设线性可分的样本集有 n 个样本 (x_i, y_i) , 其中 $i = 1, 2, \dots, n, x \in R^d, y \in \{-1, 1\}$ 是类别标号。在高维空间中,将两类样本无错分开的分类超平面满足

^{*}) 国家自然科学基金(60275002)和国家 863 计划(2003AA133060)资助。

$$g(x) = w \cdot x + b = 0 \quad (5)$$

通过对向量系数 w 进行归一化,可以使所有样本满足 $|g(x)| \geq 1$,这样分类间隔就等于 $2/\|w\|$,因此使分类间隔最大实际上就是使 $\|w\|$ 最小;而要求分类线对所有样本正确分类,就要满足

$$y_i[(w \cdot x_i) + b] - 1 \geq 0, i = 1, 2, \dots, n \quad (6)$$

根据上述讨论,最优分类面问题可表示成如下的约束优化问题,即在式(6)的约束下,求函数

$$\Phi(w) = \frac{1}{2} \|w\|^2 = \frac{1}{2} (w \cdot w) \quad (7)$$

的最小值。为此,可利用 Lagrange 函数把原问题转化为较简单的对偶问题:在约束条件

$$\sum_{i=1}^n y_i \alpha_i = 0 \quad (8)$$

$$\alpha_i \geq 0, i = 1, 2, \dots, n \quad (9)$$

之下对 α_i 求解下列函数的最大值:

$$Q(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \quad (10)$$

求解上述问题后得到的最优分类函数是:

$$\begin{aligned} f(x) &= \text{sgn}\{(w^* \cdot x_i) + b\} \\ &= \text{sgn}\left\{\sum_{i=1}^n \alpha_i^* y_i (x_i \cdot x) + b^*\right\} \end{aligned} \quad (11)$$

常数 b^* 是分类的域值,可以根据(11)式,由两类样本中任意一个支持向量求得。

考虑到一般的分类问题在定义的特征空间中并不一定线性可分,因此考虑通过一个非线性映射 T 将特征 x 映射到高维线性特征空间 F 中。高维线性空间中的内积可以定义为:

$$K(x_i, x_j) = T(x_i) \cdot T(x_j) \quad (12)$$

$K(x_i, x_j)$ 称为核函数,它的选择需要满足 Mercer 条件。采用不同的内积核函数将导致不同的支持向量机算法,目前采用较多的 3 类核函数包括:

(1) 多项式内积函数:

$$K(x, y) = [(x \cdot y) + 1]^q$$

(2) 径向基函数:

$$K(x, y) = \exp\left(-\frac{\|x - y\|^2}{\sigma}\right)$$

(3) S 型内积函数:

$$K(x, y) = \tanh(v(x \cdot y) + c)$$

支持向量机方法的优点在于没有必要知道映射 T 的具体形式,而只需定义高维空间中的内积运算即可,这样即使变换后空间维数增加很多,计算的复杂度也没有太大的变化。

3 实验结果

本实验中所用到的恒星光谱数据来自天文数据中心(Astronomical Data Center, ADC)。实验中采用了 Jacoby (1984) 中的 161 条光谱数据。按照温度由高到低,一共有七种主要的恒星光谱数据,它们

分别是 O - He II 谱线、B - He I 谱线、A - H 谱线、F - Ca II 谱线、G - strong 金属线、K - bands 扩散线和 M - very 红外线。从图 1 中我们可以看到归一化后的七类恒星谱线的分布。

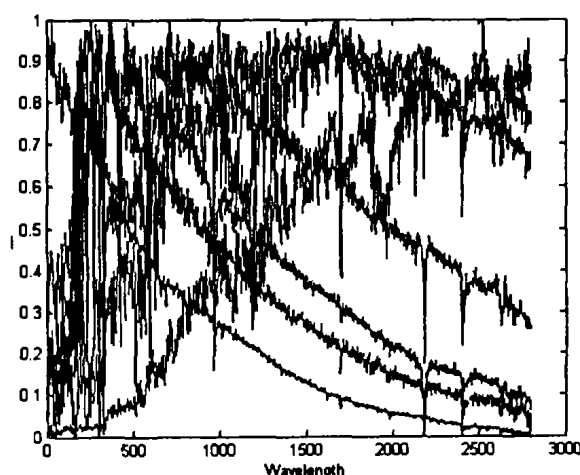


图 1 七类主要的恒星谱线

我们采用 bootstrap^[4] 技术进行实验,每类中随机选取 10 个样本作为训练样本,剩下的作为检验样本用于计算分类正确率(Correct Classification Rate - CCR)。实验重复 25 次计算 CCR 的均值和方差。

表 1 三种预处理方法的分类效果比较

	原始数据	PCA	wavelet
CCR	81.66%	81.30%	93.26%
标准差	3.75	2.90	3.08

首先,我们将原始数据直接作为 SVMs 的输入进行分类,从表 1 可以看到,分类正确率可达到 81.66%,标准差为 3.75。

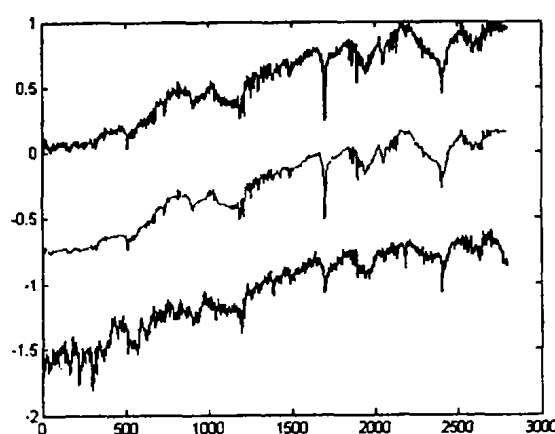


图 2 去噪效果比较:最上面的谱线是原始光谱,中间的谱线是经过小波去噪的光谱,下面的谱线是 PCA 降维重建后的光谱

考虑到光谱数据通常含有较高的噪声,所以在接下来的实验中我们采用小波变换的方法对原始数据进行降噪,降噪后的数据作为输入送入 SVMs 进行分类。从图 2 中可以看出小波去噪的效果,实验

(下转第 65 页)

似图像灰度级明显区别于其它图像的情况下适用。

讨论 本文通过在两种特征空间中对 8 种相似度量度的检索效率比较,发现两种以前较少用于图像检索的度量方法在对遥感图像的直方图特征匹配中表现出色,它们分别是 χ^2 统计距离度量和相似夹角余弦度量。由于实验中所抽取的特征类型比较单一,而遥感图像的特征又多种多样,故上述相似性度量在遥感图像检索中的适用性以及优缺点还有待于在各种不同的特征空间中深入研究。

检索图像到目前为止仅限于单波段遥感图像,且为检索实验构建的图像库规模也较小,一定程度上限制了实验结论的推广,这些都是需要在下一步工作中进一步探索和改进的问题。另外从数据模型的角度来看,本文所比较的这两组相似性度量均属于几何矩阵模型,而由于人眼视觉感知的特殊性,一般的几何矩阵模型并不能很好地反映人眼的视觉判断,因此人们一致在寻求更加有效的相似度量方法。

故今后的主要任务是将相似性度量的比较实验扩充到多源数据的检索,最终的研究结论将为在多源数据检索中发展新的相似性度量提供理论基础。

参考文献

- 1 Maxwell B A. Distance Measures for Heterogeneous Geographic Data Sets: [LOICZ report]. Swarthmore College, 2001
- 2 Jain R, et al. Similarity Measures for Image Databases. Bruce's LOICZ page, pp.1247 ~ 1254, 1995. 1247 ~ 1254
- 3 Swain M J, Ballard D H. Color indexing. International Journal Of Computer Vision, 1991, 7(1): 11 ~ 32
- 4 Gonzalez R C, Woods R E. Digital Image Processing. 3rd ed. Addison - Wesley, 1992
- 5 徐长勇,周焰,李德仁.基于内容的遥感图像检索综述.武汉理工大学学报信息与管理工程版, 2003, 26(5): 8 ~ 12
- 6 宁书年,等.遥感图像处理与应用.地震出版社, 1995. 25 ~ 32
- 7 王文惠.基于内容的图像检索技术研究:[国防科学技术大学研究生院工学博士学位论文]. 2001. 32 ~ 62
- 8 章毓晋.基于内容的视觉信息检索.科学出版社, 2003. 64 ~ 71

(上接第 58 页)

结果显示了分类正确率高达 93.26%, 标准差为 3.08。通过假设检验(t 检验)可以看出,在显著性水平 $\alpha = 0.05$ 的情况下,利用小波降噪结合 SVMs 的方法对恒星光谱数据进行分类的正确率要明显优于直接用 SVMs 进行分类的正确率(p -value = $1.97e-008$)。

主分量分析(Principal Component Analysis - PCA)^[5]是一种用较少数量的特征对样本进行描述以达到降低特征空间维数的方法,它广泛应用于信号处理、统计学和神经网络计算,是数据降维和去噪的有力工具。作为比较,我们首先用 PCA 对原始数据进行预处理,从图 3 中可以看出,前 10 PCs 的重建错误率只有 0.43%,所以我们将原始数据降至 10

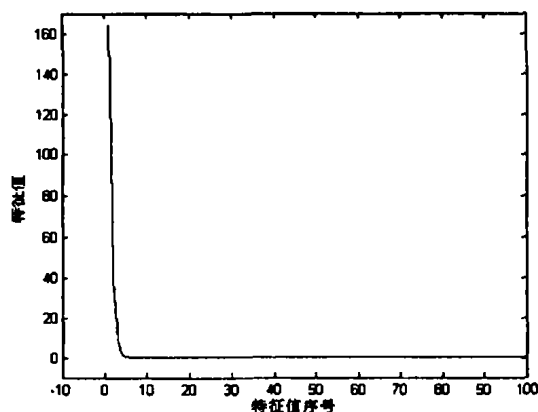


图 3 特征值降序排列

维,然后送入 SVMs 完成最终分类。从图 2 中可以看到经 PCA 降维重建后的谱线,实验结果显示了此种方法的分类正确率只有 81.30%,通过假设检验(t 检验)可以看出,在显著性水平的情况下,用 PCA

+ SVMs 进行分类的分类正确率和直接用 SVMs 分类相比较并未有显著提高(p -value = 0.71)。同时也可以看出 Wavelet + SVMs 的方法比 PCA + SVMs 的方法有更高的识别率(p -value = $2.01e-010$)。导致应用 PCA 进行降维预处理分类正确率未有显著提高的原因是由 SVMs 的基本思想决定的, SVMs 是将非线性可分样本经非线性变换映射到高维可分的空间,也就是说 SVMs 在高维样本的处理上有很好的性能。

总结与讨论 本文提出了一种恒星光谱分类的新思路,文中的组合分类器结合了小波变换和支持向量机技术。从实验结果我们可以看出,这种组合分类的方法有很好的分类效果,我们从统计学的角度用假设检验(t 检验)证明了这种组合分类器的分类效果明显优于直接用 SVMs 进行分类的效果,也优于用 PCA 结合 SVMs 的分类效果。尽管这种组合分类的方法还需要进一步的评估,但本文已经充分地显示了其准确稳定的分类性能,我们相信这种组合分类器将会被广泛地应用在恒星光谱分类中。

参考文献

- 1 Vapnik V. The Nature of Statistical Learning Theory. New York: Springer - Verlag, 1995
- 2 Cortes C, Vapnik V. Support Vector Networks. Machine Learning, 1995, 20: 273 ~ 297
- 3 Donoho D L. De - noising by soft - thresholding. IEEE Transaction on Information Theory, 1995, 41: 613 ~ 607
- 4 Efron B, Tibshirani R. An Introduction to the Bootstrap. Chaoman & Hall, 1993, 436
- 5 Jolliffe I T. Principal Component Analysis. Springer - Verlag, 1986