

一种提高恒星光谱识别率的新方法^{*}

白 凌 郭 平

(北京师范大学计算机科学系 北京 100875)

摘 要 本文提出的方法是基于 Kalman 滤波和径向基函数神经网络的,是对提高恒星光谱识别率的深入研究。它不再依赖分类正确率曲线选取特征,可以直接将 Kalman 滤波的输出作为径向基函数神经网络的输入。该技术分为两个阶段:第一阶段,采用 Kalman 滤波对经过归一化的光谱数据进行去噪、降维,同时进行第一次的分类;第二阶段,利用径向基函数神经网络进行第二次的分类。实验表明这种新技术是健壮的、高效率的、通用性强的。该方法的分类结果比主成分分析的最好结果还要好。

关键词 Kalman 滤波,径向基函数神经网络,主成分分析,恒星光谱

1 引言

在对光谱的研究当中,对光谱数据进行分类是一大难题。光谱识别通常是利用光谱线的波长、强度和谱线宽度等特征信息进行识别。识别方法一般是采用分类技术,将待识别光谱与已知的模板光谱匹配,从而将待识别光谱划分到相应的类别上。在实际中由于获取光谱时的测量误差,噪声干扰,识别技术的限制,使得待测光谱与模板很难匹配准确,造成光谱识别率较低。针对这些问题,本文主要讨论从分类技术方面去改进。

恒星光谱数据都是多维的,且是多噪声的,因此要对恒星光谱数据进行分类,所采用的方法就要能够高效地进行计算并能够有效地去除噪声。传统的多参数分类方法,如线性甄别分析(Linear Discriminant Analysis - LDA)、二次甄别分析(Quadratic Discriminant Analysis - QDA)由于数据奇异性的问题都不能直接用于这种病态数据。主成分分析(Principle Component Analysis - PCA)^[1]技术往往直接被用来对这些病态数据进行特征提取。将样本投影到低维空间上,同时尽量保留主要分量,最大限度地保留有用信息,这样提取出来的特征数就会低于样本数^[2]。另外一个可以用来解决这种奇异问题的方法就是采用 Kalman 滤波^[3]。Kalman 滤波是一种线性滤波情况下的最小均方误差估计法。它在一个独立的矩阵(光谱数据)和一个非独立的矩阵(类别标号)之间建立一个线性的模型,

$$Y = X * B + \text{Error} \quad (1)$$

而 B 矩阵是由最小均方误差估计法来估计的。

径向基函数神经网络(Radial Basis Function - RBF)是一种前馈型三层神经网络,借助其所具有的非线性模型的能力与局域化的特点,将特征光谱与

类别之间建立非线性映射关系来实现识别的功能。

2 背景

2.1 预处理技术

实验所采用的数据是从天文数据中心(Astronomical Data Centre)中选取的。其中包括 Jacoby(1984)中的 161 条光谱数据和 Pickles(1985)中的 96 条光谱数据,它们的分辨率分别是 0.14nm 和 1.2nm。这三组数据经过线性插值统一波段范围为 360 ~ 742nm,分辨率为 0.5nm。

由于所选数据来自两个库中,观察的时间、环境、噪声均不一样,因此在分类之前先对数据进行了归一化处理。

2.2 Kalman 滤波

Kalman 滤波最先是 R. E. Kalman 提出用来用于处理卫星轨道数据问题的^[4]。

在式 1 中,我们建立了一个线性模型。其中 $X_{m \times n}$ 是一个有 m 个物体和 n 维参数的光谱数据样本,而 $Y_{m \times p}$ 是一个有 m 个物体和 p 类的类别标号矩阵。 $B_{n \times p}$ 是一个有 p 列的 b - 系数矩阵。在 Kalman 滤波中, B 矩阵和它的协方差矩阵 $P_{n \times n}$ 是通过下面的公式迭代算出的^[5]。

$$B(k) = B(k-1) + K(k)[Y(k) - X(k)B(k-1)] \quad (2)$$

$$P(k) = [I - K(k)X(k)]P(k-1)[I - K(k)X(k)] + K(k)r(k)K'(k) \quad (3)$$

$$K(k) = P(k-1)X'(k)[X(k)P(k-1)X'(k) + r(k)]^{-1} \quad (4)$$

其中, n 是系统中数据的维数; m 是物体的数量; k = 1, 2, ..., m 是测量物体的索引,表明现在正在对哪个物体进行处理; B(k) 是系统状态矩阵; Y(k) 是类别标号矩阵(第 k 个物体的类别编号); X(k) 是测量矩阵(第 k 个物体的光谱); K(k) 是

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

Kalman 滤波获得的增益矩阵; $P(k)$ 是系统的协方差矩阵; $r(k)$ 是测量过程中的噪声值。

为了能够应用 Kalman 滤波, $r(k)$ 的值必须在一开始就给定。我们使用一个很小的测量值 2.5×10^{-9} 作为它的初始值。当我们给 B 矩阵和 P 矩阵赋以不同的初始值时, Kalman 滤波的结果是相对稳定的, 因为我们所建立的是一个线性的模型。在这次讨论中, 我们先用零矩阵作为 B 矩阵的初始值而单位矩阵作为 P 矩阵的初始值。对于每一个样本 k , B 矩阵、P 矩阵及向量 K 都是根据输入的第 k 个样本的数据由式(2~4)来更新的。

2.3 RBF 神经网络

RBF 神经网络由三层组成, 输入层节点只是传递输入信号到隐层, 隐层节点(亦称 RBF 节点)由激励函数构成, 而输出层节点通常是简单的线性函数。

径向基函数是以一个基函数作为一个神经元函数而以参数 W 作为权值的。

RBF 神经网络的输出为隐层节点输出的线性组合, 即^[6]:

$$Z_k(x) = \sum_{j=1}^H w_{kj} \varphi_j(\|x - \mu_j\|) + w_{k0}$$

$$= \sum_{j=0}^H w_{kj} \varphi_j(\|x - \mu_j\|) \Leftrightarrow Z(x) = W\varphi(x) \quad (5)$$

其中 w_{k0} 是偏置常数, $\varphi(\|x - \mu_j\|) = 1$, 权值矩阵 W 可以被初始化为 $W = t^{-T} \phi^{-1}$, t^{-T} 是相应的目标向量, 而 ϕ 是对称矩阵。

$$\phi = \begin{pmatrix} \varphi(\|x_1 - \mu_1\|) & \cdots & \varphi(\|x_m - \mu_1\|) \\ \vdots & \ddots & \vdots \\ \varphi(\|x_1 - \mu_p\|) & \cdots & \varphi(\|x_m - \mu_p\|) \end{pmatrix}$$

虽然有各种各样的基函数, 本文中选择的基函数 ϕ 为高斯型:

$$\varphi_j(\|x - \mu_j\|) = \exp\left[-\frac{1}{2\gamma_j^2} \|x - \mu_j\|^2\right] \quad (6)$$

其中, φ_j 是第 j 个隐层节点的输出, x 是输入样本, μ_j 是高斯函数的中心值, γ_j 是标准化常数。每个基函数的 μ_j 和 γ_j 都可以不同, 而矩阵 W 包含了权值和位移的信息。由式(6)可以看出, 节点的输出范围在 0 和 1 之间, 且输入样本愈靠近节点的中心, 输出值愈大。

3 实验

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、B、A、F、G、K 和 M。七种主要的光谱型见图 1, 具体描述见文[7]。

实验所采用的数据共有 257 个样本, 每个样本的维数均为 765, 即 $m = 257$, $n = 765$ 。其中, O 类数据共 21 条, B 类数据共 43 条, A 类数据共 33 条, F 类数据共 39 条, G 类数据共 47 条, K 类数据共 42 条, M 类数据共 32 条。我们采用交叉检验技术来

进行实验, 也就是说从每一类中随机选取 16 个样本, 每次选取其中的 15 个作为训练样本, 剩下的一个作为检验样本, 然后计算出平均的分类正确率 (Correct Classification Rate - CCR)。

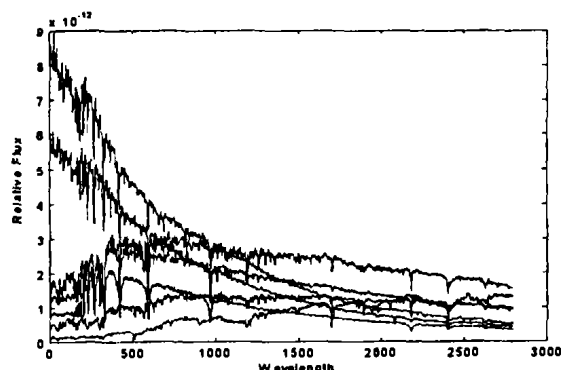
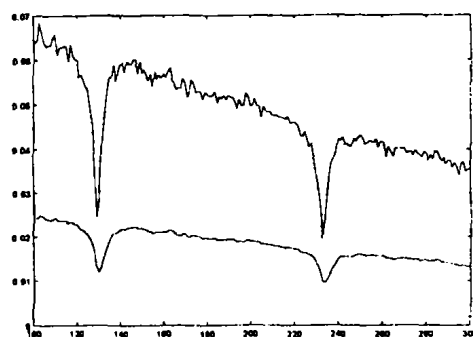


图 1 七类典型的恒星光谱

首先进行归一化处理, 实验中我们采取的办法是对每一类数据分别进行归一化, 这是因为我们认为同一类数据在实验中的表现是相似的。

然后, 采用 Kalman 滤波技术对光谱数据进行去噪降维^[5]。图 2 显示了去噪效果。从图中可以看出, 光谱曲线更加光滑, 并且保留了数据的尖峰。Kalman 滤波在一定程度上可以很好地去噪。



Kalman 滤波结果(上面一根是滤波前的光谱线, 下面的一根是滤波后的光谱线)

最后, 将 Kalman 滤波的结果 $Y = X * B$ 作为输入送入 RBF 神经网络, 进行分类。在 RBF 神经网络中, 我们使用了 21 个基函数, 选择了训练样本的 5 个不同的 γ 值。对于输出向量 z , 我们采用 one-of-k 编码方式。

由于这组数据的噪声比较大, 采用我们的新技术能达到的平均 CCR 为 0.91。

在所实验中, 开始选取零矩阵作为 B 矩阵的初始矩阵。实验证明, 当选取不同的初始值作为 B 矩阵的初始矩阵时, 分类结果都是相对稳定的, 这是因为我们建立的是一个线性模型。

RBF 网络的结果对分类结果的影响是很大的。当选取不同的神经元个数时, RBF 所得到的分类结果是很不同的。如果神经元个数选取合适, 可以大大地提高分类结果。在实验中, 我们也选取了 70 个神经元结构的 RBF 网络, 它所能达到的平均 CCR

为 0.69。当采用过多的神经元个数时,会产生过度拟合的现象。

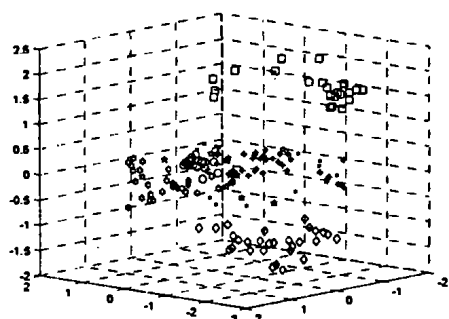


图 3 Kalman 滤波在 Y-空间的三维分布

在文[9]中,所采用的方法是将 $Y = X * B * B^T$ 作为输入送入径向基函数神经网络,然后依据 CCR 曲线选取最佳特征数。对于这组数据,它所能达到的分类正确率为 0.87。

根据式(1)中的线性模型,我们估计了 Kalman 滤波的输出。图 3 显示了它的三维输出。从图中可以看出,七类数据大致可以分开,再经过 RBF 网络进行分类,应该能得出较好的分类结果。

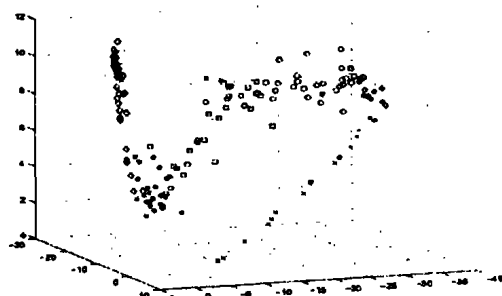


图 4 PCA 三主分量的空间分布

主成分分析(PCA)作为一种降低数据维数的技术可以直接作用于这种病态的问题。作为和 Kalman 滤波提取特征比较,本文也采用了 PCA 的方法提取出主分量^[1],对原始数据进行降维。实验表明,它所能达到的 CCR 仅为 0.71。由于 PCA 只是对线性可分的数据才能产生较好的结果,且 PCA 技术不具备滤噪的性能。

图 4 显示了 PCA 三主分量后的结果。比起 Kalman 滤波后的结果就要差很多。

表 1 给出了各种方法 CCR 的比较。

表 1 各方法 CCR 比较

方法	CCR
新方法(21 神经元)	0.91
新方法(70 神经元)	0.69
原方法(21 神经元)	0.87
PCA	0.71

结论 本文将由 Kalman 滤波和 RBF 神经网络组成的自动识别技术应用于恒星光谱的识别。我们所提出的这个新技术中,Kalman 滤波是作为一个去噪声和特征提取的预处理器,再结合 RBF 神经网络给出了相当好的结果。分类结果比使用 PCA 降维,再用 RBF 网络分类效果要好得多。

参考文献

- 1 Jolliffe I T. Principal Component Analysis. New York, Springer-Verlag, 1986
- 2 Martens H, Naes T. Multivariate Calibration. New York: Wiley, 1991
- 3 Rutan S C. Fast On-line Digital Filtering. Chemometrics Intell. Lab. Systems, 1989, 6:191 ~ 201
- 4 Rutan S C, Brown S D. Model Error Compensation in Multi-Component Analysis Using Adaptive Kalman Filtering. Anal. Chim. Acta, 1984, 160:99 ~ 119
- 5 Wu W, Rutan S C, Baldwin A. Feature Selection using the Kalman Filter for Classification of Multivariate Data. Analytica Chimica Acta, 1996, 335:11 ~ 22
- 6 Bishop C M. Neural Networks for Pattern Recognition. Oxford: Oxford University Press, 1995
- 7 Qin D M, Guo P, Hu Z Y, Zhao Y H. Automated Separation of Stars and Normal Galaxies Based on Statistical Mixture Modeling with RBF Neural Networks. Chin. J. Astron. Astrophys, 2003, 3:277 ~ 286
- 8 Bai Ling, Li ZhenBo, Guo Ping. Classification of Stellar Spectral Data Based on Kalman Filter and RBF Neural Networks. In: IEEE Int Conf. on System, Man and Cybernetics, Washington DC, 2003.274 ~ 279