第56卷 第1期 天 文 学 报 Vol.56 No.1
2015年1月 ACTA ASTRONOMICA SINICA Jan., 2015

doi: 10.15940/j.cnki.0001-5245.2015.01.004

多任务 Lasso 回归法在恒星光谱物理参量估计中的应用*

常丽娜 张培爱†

(暨南大学信息科学技术学院 广州 510632)

摘要 多任务学习方法在机器学习、计算机视觉、人工智能领域已得到广泛关注, 利用任务间的相关性, 将多个任务同时学习的效果优于每个任务单独学习的情况. 采用多任务 Lasso 回归法 (Multi-task Lasso Regression)用于恒星光谱物理参量的估计, 不仅可以获取不同物理参量间的共同的特征信息, 而且也可以很好地保留不同物理参量的特有的补充信息. 使用恒星大气模拟模型合成光谱库 ELODIE 中的光谱数据和美国大型巡天项目 Sloan 发布的 SDSS 实测光谱数据进行实验, 模型估算精度优于相关文献中的方法, 特别是对重力加速度 (lg g) 和化学丰度 ([Fe/H]) 的估计. 实验中通过改变光谱的分辨率, 施加不同信噪比 (SNR) 的噪声, 来说明模型的稳定性强. 结果表明, 模型精度受光谱分辨率和噪声的影响, 但噪声对其影响更大, 可见, 多任务 Lasso 回归法不仅操作简便, 稳定性强, 而且也提高了模型的整体预测精度.

关键词 恒星: 基本参数, 方法: 数据分析, 方法: 统计, 方法: 其它诸多方面

中图分类号: P144; **文献标识码**: A

1 引言

现在的大口径兼大视场望远镜, 如我国的郭守敬望远镜 [1−2] , 美国的 Sloan 数字巡天望远镜 [3−4] 等, 可以得到大量的光谱数据. 如何在海量光谱数据中测量出恒星光谱的物理参量也成为了天体光谱数据处理中最基本、最重要的内容. 恒星光谱物理参量主要有表面有效温度 ($T_{\rm eff}$) 、重力加速度 (lg g) 与化学丰度 ([Fe/H]) . 目前, 人们研究提出了多种关于恒星光谱物理参量自动估计的方法, 主要有人工神经网络 (Artificial Neural Network, ANN) 算法和最近邻算法 (Minimum Distance Method, MDM). 如由 Bailer-Jones 设计的 820: 5: 5: 1 结构的 ANN [5] 用于恒星表面有效温度的预测, 2000年又开发了双隐层、多感知器的前馈 ANN 系统; Fuentes 等的 K-近邻算法 [6]、Allende 的加权平均算法 [7]、Zhang 等的变窗宽非参数回归法 [8−9] 等都是 MDM 的变形.

2014-05-16 收到原稿, 2014-08-06 收到修改稿

*教育部人文社会科学研究一般项目 (11YJAZH118) 资助

†qzhzhang@163.com

由于光谱数据海量的特点, 以及光谱在传输、接收过程中往往受到大量噪声的干扰, 例如光子噪声、天光线以及设备的噪声等, 影响了最终恒星光谱物理参量的估计效果. 在进行模型预测前, 要对高维数据降维、剔除噪声干扰, 相关的方法有主成分分析 (PCA) 、滤波法等. Tibshirani 于1996年提出的 Lasso [10] (Least Absolute Shrinkage Selection and Operator) 算法由于其计算速度快、精度高, 备受青睐. 恒星光谱包含了恒星中物理参量的信息, 但目前许多模型分开考虑恒星光谱物理参量, 失去了物理参量间潜在的联系. 近年来, 在机器学习、计算机视觉、人工智能领域, 多任务学习 (Multi-task Learning, MTL) 方法引起了众多学者的研究兴趣, 目的是获取不同任务间的潜在关系, 将多个相关任务同时学习, 进而充分利用任务间丰富的信息. 这样的学习方法有利于任务的互相学习, 更能突显它能提高预测模型的预测效果和泛化性能的优势. 比如, Evgeniou 等 [11] 使用多任务支持向量机的方法用于提高消费者消费偏好的预测准确率; Bakker 等 [12] 通过实验说明了在少量图像类别情况下多任务分类方法能够提高分类器的泛化性能. 虽然使用 Lasso 可以对每个任务独立地进行有效的学习 [13], 但它忽略了任务间潜在的联系, 易造成数据的过度拟合, 影响最终模型的表现效果. 在多任务学习中, Liu 等 [14] 通过有效的 L_{21} 范式最小化可以让不同的任务获取共同的特征, 然而这种约束过强, 没有考虑每个任务所特有的特征信息, 可能会降低预测模型的泛化性能. 而本文使用的多任务 Lasso 可以克服这些不足, 在对多个任务同时学习的同时, 不仅可以获取不同任务间的共同的特征信息, 而且也可以很好地保留不同任务的补充信息 [15]. 如果将建模分析每个物理参量的问题看作一个任务, 则可将物理参量同时建模分析的问题转换为多任务学习问题, 从真正意义上实现物理参量的同时建模分析. 然后在多任务 Lasso 的基础上进行通常的线性回归对恒星光谱物理参量 $\lg T_{\rm eff}$、$\lg g$、[Fe/H] 进行估计, 它避免了多个任务分开独立计算的繁琐, 更重要的是提高了恒星光谱物理参量的整体估计精度和预测模型的泛化性能.

2　多任务 Lasso 回归法

在大数据时代的背景下, 多任务处理变得尤为重要.假设有 m 个任务, 给定数据 $\boldsymbol{X}^j = (\,\boldsymbol{x}_1^j, \cdots, \boldsymbol{x}_i^j, \cdots, \boldsymbol{x}_n^j\,)^{\rm T} \in \boldsymbol{R}^{n\times d}$, n 是样本数, d 是样本特征变量数, $j = 1, \cdots, m$. $\boldsymbol{X}^j$ 所对应的响应变量 $\boldsymbol{y}^j = (\,\boldsymbol{y}_1^j, \cdots, \boldsymbol{y}_i^j, \cdots, \boldsymbol{y}_n^j\,)^{\rm T} \in \boldsymbol{R}^{n\times 1}$. 对于恒星光谱, 涉及的所有物理参量对应的光谱数据是一样的, 即对所有任务, 输入样本 $\boldsymbol{X}^j$ 是相同的, 但本文使用的方法不限于此, 它具有更广的适用性. 需要预测的线性回归模型 [16] 为:

$$\hat{\boldsymbol{y}}^j = \boldsymbol{X}^j \boldsymbol{w}^j\,, \tag{1}$$

其中, $\boldsymbol{w}^j \in R^{n\times 1}$, 表示任务 j 的回归系数向量. 为了同时计算 m 个任务的 m 个回归系数向量, 即 $\boldsymbol{W} = (\,\boldsymbol{w}^1, \cdots, \boldsymbol{w}^j, \cdots, \boldsymbol{w}^m\,)$, 需要优化的多任务 Lasso 模型为:

$$\mathrm{Min}_{\boldsymbol{W}} \sum_{j=1}^{m} ||\boldsymbol{X}^j \boldsymbol{w}^j - \boldsymbol{y}^j||_{\rm F}^2 + \lambda_1 ||\boldsymbol{W}||_1 + \lambda_2 \boldsymbol{D}\,, \tag{2}$$

其中, $\boldsymbol{D} = \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1,k\neq j}^{m} \frac{||\boldsymbol{x}_i^j \boldsymbol{w}^j - \boldsymbol{x}_i^k \boldsymbol{w}^k||_{\rm F}^2}{||\boldsymbol{x}_i^j - \boldsymbol{x}_i^k||_{\rm F}^2}$. $||\cdot||_1$ 表示 1 - 范式, $||\cdot||_{\rm F}$ 表示 Frobenius

- 范式, 正则化参数 λ_1 控制着模型的稀疏性, 当 λ_1 增大时, 模型 $\boldsymbol{W}$ 的稀疏性逐渐增强, 即 $\boldsymbol{W}$ 中的非零元素逐渐减少, λ_2 控制着不同任务的信息保留程度.

虽然传统的 Lasso 使用稀疏回归 (基于 L_1 - 范式) 可以有效、独立地对每个任务进行预测, 但它忽略了任务间潜在的联系, 易造成数据的过度拟合. 基于 L_{21} 范式的组稀疏可以让不同的任务获取共同的特征, 然而这种约束过强, 没有考虑每个任务所特有的特征信息, 可能会影响预测模型的泛化性能. 上述 (2) 式中, 不同任务在获取共同的特征信息的同时, 正则化项 $\boldsymbol{D}$ 又能有效地保留不同任务的特有补充信息.

对于 (2) 式的求解采用加速梯度法 [17] (Accelerated Gradient Method, AGM), AGM 不像传统的梯度法, 在每次迭代中只用最近的点作为当前的搜索点, 而是用前两个点的一个线性组合作为新的搜索点, 使收敛速度更快. 同时, 为了确定参数 λ_1 、 λ_2 , 我们使用交叉验证 (Cross Validation) 进行参数优选.

3 数据

实验数据一: 选取恒星大气模拟模型合成光谱库 ELODIE 中的 1 800 条光谱数据用于实验, 所有的光谱均已经过流量校准. 所有样本的光谱波长 $\lambda = 421 \sim 650$ nm, 光谱的分辨率 $\Delta\lambda = 1$ nm. 3 个物理参量的数据范围分别为: $T_{\rm eff}$: 3700 ~ 13386 K, lg g: 0.00 ~ 4.80 dex, [Fe/H]: $-2.94 \sim 1.00$ dex .

实验数据二: 选取美国大型巡天项目 Sloan 发布的 SDSS-DR7 中的 4 000 条恒星光谱数据. 这些光谱来自 102 个板块 (0266—0367), 每个板块最多可观测到 640 条光谱. 实际中随机选用每个板块的部分恒星光谱数据用于实验, 在对数波长格式下将其移动到静止波长, 截取共同波长 $\lambda = 398 \sim 794$ nm, 并使用线性插值按照分辨率 $\Delta\lambda = 0.1$ nm 对光谱进行采样. 3 个物理参量的数据范围分别为: $T_{\rm eff}$: 4163 ~ 9685 K , lg g: 1.26 ~ 4.99 dex , [Fe/H]: $-3.44 \sim 0.18$ dex .

为了更精确地对温度进行描述, 实验中用温度的对数值 lg $T_{\rm eff}$ 代替温度 $T_{\rm eff}$. 对每个物理参量的测量效果, 采用平均绝对误差 δ (mean absolute error: δ)、误差的标准差 v (standard deviation: v)和平均误差 u (mean error: u) 来度量.

4 实验结果与分析

基于 ELODIE 合成光谱数据, 在实验中随机选取 ELODIE 合成光谱库中的 1 800 条光谱, 分成两部分, 75% 的样本作为训练集, 剩下 25% 的样本作为测试集. 每条光谱在训练和测试之前, 首先进行二范数行归一化的预处理, 归一化操作为: 已知 n 条 d 维的光谱数据 $\boldsymbol{X}_{ip}^{j'}$, $\boldsymbol{X}_{ip}^{j} = \boldsymbol{X}_{ip}^{j'} / \sqrt{\sum_{p=1}^{d} (\boldsymbol{X}_{ip}^{j'})^2}$ ($i = 1, 2, \cdots, n; p = 1, 2, \cdots, d$), 然后用多任务 Lasso 回归法对恒星光谱物理参量做估计. 我们把这种方法同文献 [18-19] 的方法做对比, 有基于主成分分析的非参数回归法 (PCA+non-parameter)、基于 Haar 小波的非参数回归法 (Haar+non-parameter)、基于主成分分析的支持向量机回归法 (PCA+SVR)、基于 Haar 小波的支持向量机回归法 (Haar+SVR). 参量 lg $T_{\rm eff}$ 、 lg g 、[Fe/H] 平均绝对误差 δ 和误差的标准差 v 的统计结果见表 1 .

表 1 多任务 Lasso 回归法和相关文献方法对 ELODIE 数据 lg $T_{\rm eff}$、lg g、[Fe/H] 的预测结果误差的比较

Table 1 The error comparison of the predicted lg $T_{\rm eff}$, lg g, and [Fe/H] in the ELODIE data estimated with the multi-task Lasso regression and the methods in the related literature

Method \ Err.	$\delta_{\lg T_{\rm eff}}$	$\upsilon_{\lg T_{\rm eff}}$	$\delta_{\lg g}$	$\upsilon_{\lg g}$	$\delta_{\rm [Fe/H]}$	$\upsilon_{\rm [Fe/H]}$
PCA+non-parameter	0.0148	0.0286	0.2183	0.3718	0.1749	0.3169
Haar+non-parameter	0.0174	0.0304	0.2625	0.4221	0.2422	0.3838
PCA+SVR	0.0188	0.0322	0.2193	0.3557	0.1707	0.3321
Haar+SVR	0.0204	0.0327	0.2474	0.3952	0.1992	0.3647
This method	0.0135	0.0210	0.1683	0.2350	0.1368	0.2367

由表 1 可见，多任务 Lasso 回归方法对恒星光谱物理参量的预测效果优于相关文献中的方法, 尤其是对 lg g 和 [Fe/H] 的预测. 观察表 2 , 3 个物理参量 lg $T_{\rm eff}$、lg g、[Fe/H] 的平均误差均在 0 附近, 说明系统偏差较小; 且 lg $T_{\rm eff}$、lg g、[Fe/H] 的预测值与真值之间的相关系数分别达到 0.985 3、0.966 1、0.779 8 ; 计算不同物理参量间的相关性, lg g 的残差与 [Fe/H] 残差的相关性为 0.256 9 , $T_{\rm eff}$ 残差与 [Fe/H] 残差的相关性为 0.218 9 , 但是 $T_{\rm eff}$ 的残差与 lg g 残差的相关性仅为 0.181 5 , 可见, 3 个物理量间存在相关性, 但并不是很强, 这可能与物理参量本身的性质有关, 或是恒星的演化影响了彼此间的相关性. 图 1 对上述情况进行了直观描述, lg $T_{\rm eff}$、lg g、[Fe/H] 的估计值有着很好的拟合效果, 其中 lg $T_{\rm eff}$ 的样本点分布均匀, 只有极少数偏离真实值较远, 拟合效果最好; 误差的正态分布图进一步说明了多任务 Lasso 回归模型适合进行恒星光谱物理参量的估计. 对于 [Fe/H] 的估计值与真值的对比图, 可以发现有若干偏离真值较远的点, 造成的原因有: (1) 这些点可能来自不同的星体, 而不同星体之间属性差别较大; (2) 观测或仪器的偶然因素造成部分数据偏离真值较大; (3) [Fe/H] 本身的复杂性影响了其性能的规律性表现.

表 2　多任务 Lasso 回归法对 ELODIE 数据 lg $T_{\rm eff}$、lg g、[Fe/H] 的预测结果

Table 2 The predicted results of lg $T_{\rm eff}$, lg g, and [Fe/H] in the ELODIE data estimated with the multi-task Lasso regression

Parameter \ Result	δ	v	u	R
lg $T_{\rm eff}$	0.0135	0.0210	2.66×10^{-6}	0.9853
lg g	0.1683	0.2350	-7.25×10^{-6}	0.9661
[Fe/H]	0.1368	0.2367	4.38×10^{-6}	0.7798

鲁棒性也是衡量模型优越性的因素之一, 为此将所有光谱的分辨率从 $\Delta\lambda = 1$ nm 分

别变化到2 nm、3 nm, 并分别对这些光谱添加信噪比 (Signal to Noise Ratio, SNR) 依次为 SNR = 20, 50, 100, 150, 200 的噪声. 图 2 描述了平均绝对误差 δ 在 $\Delta\lambda = 1$ nm , 2 nm , 3 nm 下随不同信噪比 SNR 的变化趋势. 3 个物理参量的预测精度受光谱分辨率、噪声的影响, 当分辨率从 $\Delta\lambda = 1$ nm 变化到 $\Delta\lambda = 3$ nm , 两种误差结果随之增大, 预测精度降低. 信噪比越低, 对 3 个物理参量的估计结果影响越大, 随着信噪比的增大, 误差逐渐减小, 当 SNR = 100 时, 误差基本趋于稳定. 综合来讲, 噪声对 3 个物理参量预测效果的影响大于分辨率对它们的影响. 误差的标准差在天文学中又叫误差的弥散度, 对比 3 个物理参量的误差的标准差, lg $T_{\rm eff}$ 的 v 值相对最小, 也说明了模型在预测 lg $T_{\rm eff}$ 时稳定性最强, [Fe/H] 次之, lg g 相对最差.

针对恒星光谱物理参量的估计, 本文又作了进一步的实验. 由于根据恒星温度的不同, 可以将恒星光谱分为 7 大类: O: $>$ 25000 K; B: 11000 $\sim$ 25000 K; A: 7500 $\sim$ 11000 K; F: 6000 $\sim$ 7500 K; G: 5000 $\sim$ 6000 K; K: 3500 $\sim$ 5000 K; M: $<$ 3500 K, 则计算实验所用 ELODIE 光谱数据的不同光谱类型的物理参量的平均绝对误差 δ, 实验结果见图 3. 其中对 F 类恒星光谱, 即有效温度 $T_{\rm eff}$ 在 6000 $\sim$ 7500 K 时, 物理参量误差较大, 但仍在误差允许的范围内.

为更好地说明该模型的有效性, 下面将其应用于 SDSS 实测光谱数据. 随机选取 SDSS 实测光谱库中的 4 000 条光谱, 分成两部分, 75% 的样本作为训练集, 剩下 25% 的样本作为测试集. 每条光谱在训练和测试之前, 首先进行二范数行归一化的预处理, 然后用多任务 Lasso 回归法对 3 个物理参量做估计. 表 3 描述了所有光谱的分辨率从 $\Delta\lambda =$ 0.1 nm, 分别变化到 $\Delta\lambda = 0.2$ nm, $\Delta\lambda = 0.3$ nm 的 3 种误差结果. 图 4 对不同类型恒星光谱物理参量的平均绝对误差进行了描述. 可见, SDSS 实测数据中 3 个物理参量的预测效果要比 ELODIE 合成数据的预测效果好, 但是有一些共同点: lg $T_{\rm eff}$ 的精度最高, [Fe/H] 次之, lg g 相对最差; 平均误差 u 都在 0 附近, 说明系统偏差小. 在不改变光谱分辨率的情况下, lg $T_{\rm eff}$、lg g、[Fe/H] 的预测值与真实值的相关系数 R 分别为: 0.991 7, 0.893 6, 0.959 9. 以上情况也说明了多任务 Lasso 回归法针对不同的数据集, 对恒星光谱物理参量的估计是稳定的, 预测模型的泛化性能比较好. 另一方面, 类似于 ELODIE 数据, 不同物理参量间存在相关性, 但不是很强, lg g 的残差与 [Fe/H] 残差的相关性为 0.310 6, $T_{\rm eff}$ 残差与 [Fe/H] 残差的相关性为 0.233 2, $T_{\rm eff}$ 的残差与 lg g 残差的相关性为 0.260 6. 这也是大样本巡天光谱数据自动分析面临的问题, 只有考虑了影响光谱的各种因素和演化模型, 大样本恒星光谱物理参量的估计才能完全自动化.

表 3 SDSS 数据物理参量 lg $T_{\rm eff}$、lg g、[Fe/H] 在不同分辨率下的实验结果

Table 3 The error analysis of lg $T_{\rm eff}$, lg g, and [Fe/H] with different resolutions in the physical parameters of SDSS data

Err. / $\Delta\lambda$	$\delta_{\lg T_{\rm eff}}$	$v_{\lg T_{\rm eff}}$	$u_{\lg T_{\rm eff}}$	$\delta_{\lg g}$	$v_{\lg g}$	$u_{\lg g}$	$\delta_{\rm [Fe/H]}$	$v_{\rm [Fe/H]}$	$u_{\rm [Fe/H]}$
0.1 nm	0.0065	0.0096	0.0004	0.1793	0.2478	0.0002	0.1193	0.1589	0.0001
0.2 nm	0.0064	0.0095	0.0001	0.1967	0.2659	0.0005	0.1335	0.1738	0.0002
0.3 nm	0.0067	0.0097	0.0002	0.2153	0.2875	0.0001	0.1498	0.1929	-0.0005

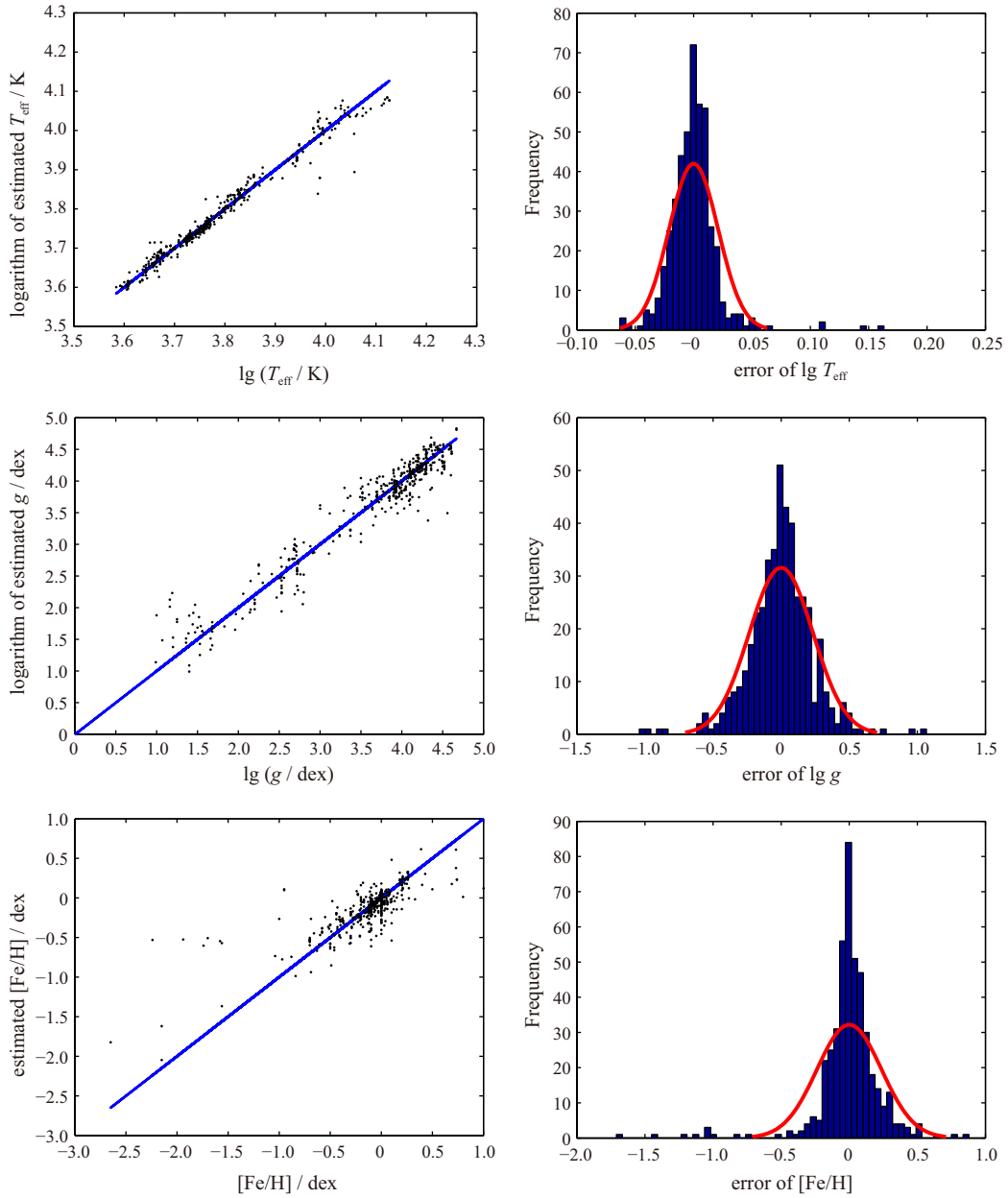

图 1　左图是光谱物理参量 lg $T_{\rm eff}$、lg g、[Fe/H] 的估计值与 ELODIE 真实值的对比; 右图是 lg $T_{\rm eff}$、lg g、[Fe/H] 的误差直方图及正态分布情况.

Fig. 1　Left: the comparison of the estimated spectral physical parameters lg $T_{\rm eff}$, lg g, and [Fe/H] in the ELODIE data with their real values. Right: the histogram and normal distribution of lg $T_{\rm eff}$, lg g, and [Fe/H] residuals

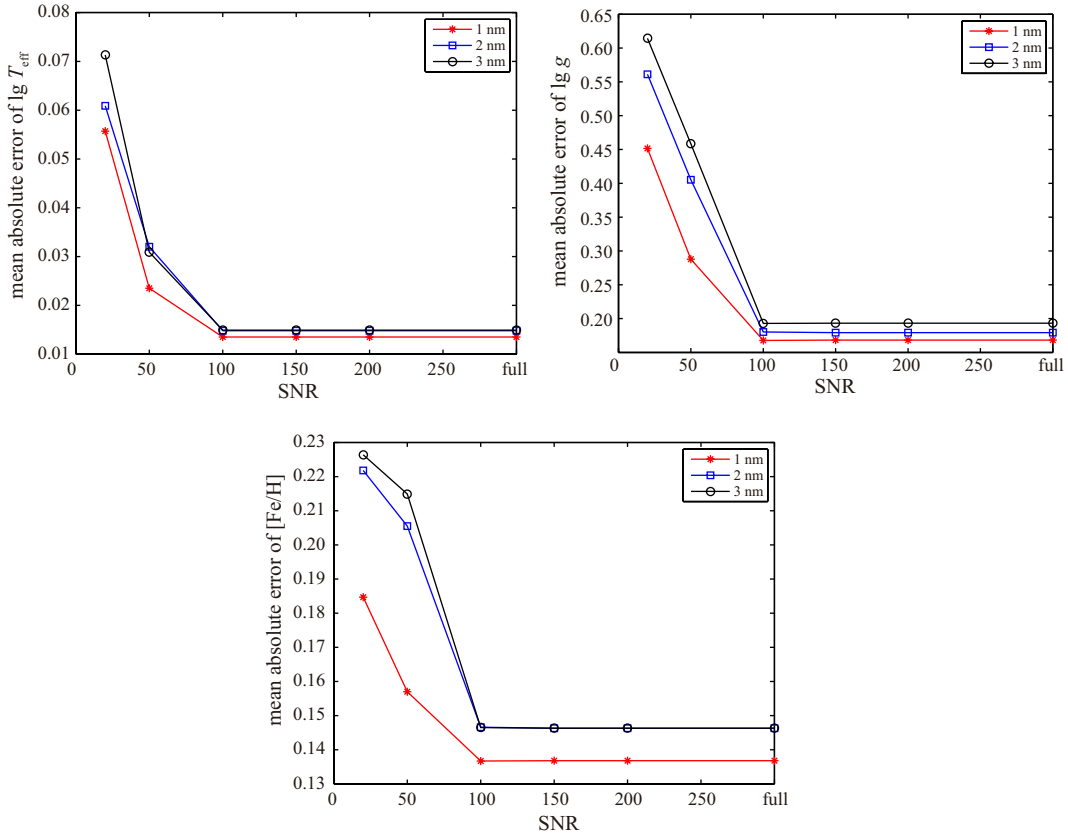

图 2 平均绝对误差在分辨率 $\Delta\lambda = 1$ nm, 2 nm, 3 nm、信噪比 SNR = 20, 50, 100, 150, 200和无噪声下的曲线图

Fig. 2 The mean absolute error curves with the SNR of 20, 50, 100, 150, 200, and full, and the resolution of $\Delta\lambda = 1$ nm, 2 nm, and 3 nm, respectively

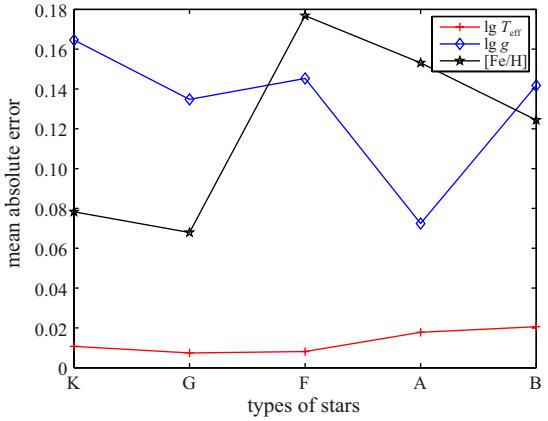

图 3 ELODIE 数据中不同类型恒星光谱的 lg $T_{\rm eff}$、lg g、[Fe/H] 的平均绝对误差曲线图

Fig. 3 The mean absolute error curves of lg $T_{\rm eff}$, lg g, and [Fe/H] in the ELODIE data for different types of stellar spectra

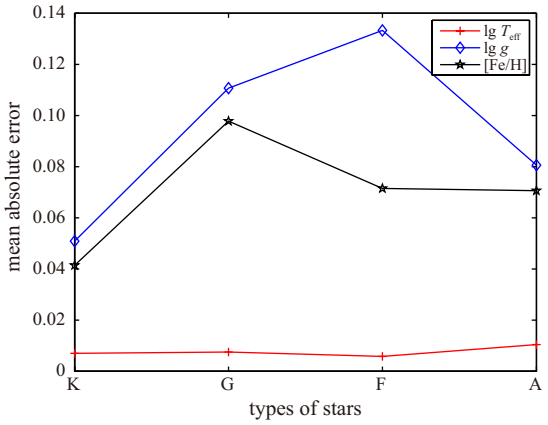

图 4　SDSS 数据中不同类型恒星光谱的 lg $T_{\rm eff}$、lg g、[Fe/H] 的平均绝对误差曲线图

Fig. 4　The mean absolute error curves of lg $T_{\rm eff}$, lg g, and [Fe/H] in the SDSS data for different types of stellar spectra

5　结论

文章采用的多任务 Lasso 回归法, 将多个物理参量同时建模估计的问题转化为 MTL 的问题, 从而可充分利用各物理参量间潜在的信息, 从真正意义上实现了物理参量的同时建模分析, 避免了繁琐的计算, 且从整体上提高了模型的预测精度和泛化性能. 使用多任务 Lasso 回归法对恒星光谱物理参量进行估计, 预测精度优于相关文献方法的预测结果, 尤其是对 lg g 和 [Fe/H] 的估计. 基于不同的光谱数据库 ELODIE 合成数据库和 SDSS 实测数据库进行实验, 说明该模型对恒星光谱物理参量进行估计的有效性. 为验证模型的稳定性, 实验中改变光谱的分辨率, 施加不同信噪比的噪声, 结果表明, 模型精度受分辨率和噪声的影响, 但噪声对其影响更大. 综上考虑, 多任务 Lasso 回归法操作简便, 稳定性强, 估算精度高, 但其需要每个任务所对应的样本特征数目相同, 所以还有待继续研究去突破这一限制, 以能将其应用于更广泛的领域.

参 考 文 献

[1] Cui X Q, Zhao Y H, Chu Y Q, et al. RAA, 2012, 12: 1197
[2] Zhao G, Zhao Y H, Chu Y Q, et al. RAA, 2012, 12: 723
[3] Noterdaeme P, Petitjean P, Carithers W C, et al. A&A, 2012, 547: L1
[4] Paris I, Petitjean P, Aubourg E, et al. A&A, 2012, 548: A66
[5] Bailer-Jones C A L. A&A, 2000, 357: 197
[6] Fuentes O, Gulati R K. RMxAC, 2001, 10: 209
[7] Allende P C. AN, 2004, 325: 604
[8] 张健楠, 吴福朝, 罗阿理, 等. 天文学报, 2005, 46: 406
[9] Zhang J N, Wu F C, Luo A L, et al. ChA&A, 2006, 30: 176
[10] Tibshirani R. JSTOR, 1996, 58: 267
[11] Evgeniou T, Pontil M. ACM, 2004: 109
[12] Bakker B, Heskes T. JMLR, 2003, 4: 83
[13] Huang T, Gong H P, Yang C, et al. CBAC, 2012, 43: 46

[14] Liu J, Ji S, Ye J. Proceedings of the 25th Conference on Uncertainty in Artifical Intelligence. Arlington: AUAI Press, 2009: 339
[15] Liu F, Chong Y W, Chen H F, et al. NeuroImage, 2013, 84: 466
[16] Zhou J, Yuan L, Liu J, et al. ACM, 2011: 814
[17] Nesterov Y. Gradient Methods for Minimizing Composite Objective Function. CORE Discussion Paper 2007/76 September 2007
[18] 张健楠, 吴福朝, 罗阿理, 等. 光谱学与光谱分析, 2009, 29: 1131
[19] 卢瑜, 李乡儒, 王永俊, 等. 光谱学与光谱分析, 2013, 33: 2010

Application of Multi-task Lasso Regression in the Stellar Parametrization

CHANG Li-na ZHANG Pei-ai
(College of Information Science and Technology, Jinan University, Guangzhou 510632)

Abstract The multi-task learning approaches have attracted the increasing attention in the fields of machine learning, computer vision, and artificial intelligence. By utilizing the correlations in tasks, learning multiple related tasks simultaneously is better than learning each task independently. An efficient multi-task Lasso (Least Absolute Shrinkage Selection and Operator) regression algorithm is proposed in this paper to estimate the physical parameters of stellar spectra. It not only makes different physical parameters share the common features, but also can effectively preserve their own peculiar features. Experiments were done based on the ELODIE data simulated with the stellar atmospheric simulation model, and on the SDSS data released by the American large survey Sloan. The precision of the model is better than those of the methods in the related literature, especially for the acceleration of gravity (lg g) and the chemical abundance ([Fe/H]). In the experiments, we changed the resolution of the spectrum, and applied the noises with different signal-to-noise ratio (SNR) to the spectrum, so as to illustrate the stability of the model. The results show that the model is influenced by both the resolution and the noise. But the influence of the noise is larger than that of the resolution. In general, the multi-task Lasso regression algorithm is easy to operate, has a strong stability, and also can improve the overall accuracy of the model.

Key words stars: fundamental parameters, methods: data analysis, methods: statistical, methods: miscellaneous