第60卷 第2期 天 文 学 报 Vol.60 No.2
2019年3月 ACTA ASTRONOMICA SINICA Mar., 2019

doi: 10.15940/j.cnki.0001-5245.2019.02.011

基于深度学习的LAMOST光谱分类研究*

许婷婷[1†] 马晨晔[2] 张静敏[1] 周卫红[1,3‡]

(1 云南民族大学数学与计算机科学学院 昆明 650500)
(2 云南农业职业技术学院经济管理学院 昆明 650031)
(3 中国科学院天体结构与演化重点实验室 昆明 650011)

摘要 光谱分类不仅对理解恒星物理学有着重要意义, 而且在研究银河系整体结构和演化过程中起着至关重要的作用. 然而在相关研究中仍存在分类精度低和光谱型未知等问题, 因此提出一种新的光谱自动分类模型并将其应用在F、G和K 3种恒星光谱的分类中, 方法的基本思想是训练一个深度信念网络对光谱数据进行分层特征学习, 然后采用反向传播算法对整个模型进行微调. 从LAMOST (Large Sky Area Multi-Object Fiber Spectroscopic Telescope) Data Release 5 (DR5)中选取31667条包含F、G和K 3种恒星的光谱数据, 并在TOPCAT软件中与GAIA (Global Astrometric Interferometer for Astrophysics)数据进行交叉, 得到颜色-星等图并验证光谱数据的分布. 最后对该模型进行评估, 结果表明: 深度信念网络在综合性能上优于其他分类算法.

关键词 恒星: 基本参数, 方法: 数据分析, 方法: 统计学, 技术: 光谱分析

中图分类号: P144; **文献标识码**: A

1 引言

伴随着天文观测仪器的问世, 例如在SDSS (Sloan Digital Sky Survey)[1]、GAIA (Global Astrometric Interferometer for Astrophysics)[2]和LAMOST (Large Sky Area Multi-Object Fiber Spectroscopic Telescope)[3]等巡天项目中大量光谱数据也随之产生. 在这些已获得的光谱数据中, 有很多是光谱型未知或者是现有分类可信度低的光谱数据, 从这些光谱数据中获得有价值的信息, 提高LAMOST望远镜的科学产出是非常有必要的, 因此对这些光谱数据进行分析研究是一项非常重要的工作.

天体光谱处理主要是自动地对光谱进行分析与测量, 抽取出光谱中所包含的各种物理信息, 如视向速度、光谱型和红移等, 并实现光谱的测量、证认和分类. 鉴

2018-10-17收到原稿, 2018-12-05收到修改稿

*国家自然科学基金项目(61561053), 云南民族大学研究生创新基金科研项目(2018YJCXS222), 中国科学院天体结构与演化重点实验室(OP201512)资助

†1430615747@qq.com

‡ynzwh@163.com

于LAMOST海量的光谱数据, 引入计算机程序进行自动或者半自动的分析处理显得尤为重要.

随着光谱观测在天文上的广泛开展, 学者们对光谱分类的方法及在天文上的应用进行了大量研究. 较早的有吴永东等[4]应用空间选择性滤波、多尺度形态滤波等技术对类星体光谱进行识别. 后来覃冬梅等[5−6]提出了两种快速的恒星光谱型分类方法, 一种是基于主成分分析方法利用最近邻分类器构建分类树进行光谱分类; 另一种方法是结合主成分分析方法提出一种新的基于支撑矢量机的非活动天体与活动天体的自动分类方法. Shi等[7]采用支持向量机(SVM)方法对星系的分类问题进行了研究. 赵瑞珍等[8]采用基于稀疏表示的方法进行谱线自动提取的研究.

LAMOST巡天项目投入运行之后, 与美国的SDSS巡天项目相比较, LAMOST没有配套的测光观测, 只有光谱数据, 在进行自动分类时不能借助色指数, 对分类识别增加了难度. 虽然LAMOST的Pipeline对光谱进行了初步的分类[9], 但由于多种原因一些恒星的分类识别结果还不是十分理想. 此后, Liu等[10]对LAMOST光谱的进一步分类研究, 发现由于巨星中B型以及早期的K型光谱与A型以及晚型的G型光谱非常相似导致分类困难, 尚未解决的主要问题包括巨星中的OB、K, 亚巨星支的A、G的分类精度非常低, 分类识别方法和结果仍然有待完善. 由以上分析可知, LAMOST光谱中还存在一些不能确定的类型或者分类可信度低的光谱数据, 针对这一问题, 计划将人工智能的最新成果用于光谱数据的分类识别中, 即采用深度学习的方法对天体光谱数据进行分类研究并结合天体物理理论进行描述.

深度学习概念起源于人工神经网络, 作为机器学习中的一个新领域由Hinton等人于2006年提出[11], 通过对人脑机制的模仿来解释图像、文本和语音等数据, 训练和学习类似于人脑的神经网络. 微软研究人员通过与Hinton合作, 首先将深度学习中的受限玻尔兹曼机(Restricted Boltzmann Machine, RBM)[12]和深度信念网络(Deep Belief Network, DBN)引入到语音识别模型中, 在词汇量较大的识别系统中效果具有明显的提升, 相比之前的语音识别错误率下降了30%. 由于深度学习的优势在于样本越大分类精度越高, 得益于LAMOST光谱数据的大样本优势, 有理由相信将深度学习方法应用于LAMOST光谱数据的分类会取得较好结果. 近期相关研究见Wang等[13]将深度神经网络方法应用到光谱分类研究和缺陷谱恢复的研究中, 潘濡扬等[14−15]利用深度学习方法对恒星大气物理参数自动估计研究, 因此相关研究工作值得进一步开展.

2 光谱数据和预处理

大天区面积多目标光纤光谱天文望远镜LAMOST是一架视场为5°、横卧于南北方向的中星仪式反射施密特望远镜, 也称为郭守敬望远镜, 是当今世界上光谱获取率最高的天文望远镜, 最多可同时获得4000个天体的光谱. 截止2017年12月31日, 包含先导巡天及正式巡天5 yr的LAMOST Data Release 5 (DR5)数据集正式发布, 其中包括4154个观测天区, 共发布了901万条光谱, 高质量光谱数(信噪比>10)达到了777万条, 恒星参数534万组, 是世界上最大的、有传承价值的天体光谱数据库, 为研究银河系的形成和演化提供了基础性数据.

实验中总共包含来自LAMOST DR5中的31667条光谱数据, 其中1667条F、G和K型

光谱数据的信噪比大于20, 样本数据取自文献[11], 并在SIMBAD天文数据库中证认其Morgan-Keenan (MK)类别. 另外考虑到1667条光谱数据样本规模较小, 为了体现深度信念网络大样本的优势, 我们引入同样来源于LAMOST DR5数据库的30000条样本, 所选取的数据已被LAMOST Pipeline分为F、G和K 3种光谱型, 每种类型的样本均为10000条. 同样3种类型的光谱, 与文献[11]中的样本不同之处在于后者对信噪比没有进行限制. 为了验证数据标注的准确性, 将其样本在TOPCAT中与GAIA数据进行了交叉, 得到颜色-星等图如图1所示, 其中3种颜色分别代表3种类型的光谱数据, 横坐标表示颜色, 纵坐标代表光谱数据的绝对星等值.

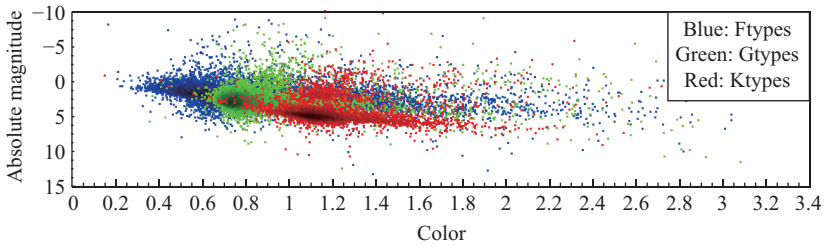

图 1　样本与GAIA数据交叉的颜色-星等图

Fig. 1　The color-magnitude diagram of sample crossed with the GAIA data

为了方便, 我们将所有光谱数据样本的维度统一调整为3909维. 但值得注意的是, 在不同波长下, 流量频谱变化很大, 即原始数据不同维度的值差异很大. 因此, 为了降低其计算复杂度且不影响光谱分类精确率, 需要对原始数据进行归一化处理, 本文采用的归一化方法是: min-max标准化, 也称为离差标准化, 是对原始数据的线性变换, 使结果映射到[0, 1]之间. 其转换函数如下:

$$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}, \tag{1}$$

其中, x_{norm}表示归一化后得到的特征值, x为光谱数据对应的每个特征值, $x_{\max}$为每条光谱样本数据的最大特征值, $x_{\min}$为每条光谱样本数据的最小特征值.

3　深度信念网络设计

3.1　受限玻尔兹曼机的结构

受限玻尔兹曼机是一种特殊的玻尔兹曼机, 具有两层结构、对称连接和无自反馈的特点, 它可用来作为基本模块构造自编码器、深层信念网络、深层玻尔兹曼机等许多其他深层模型, 其结构如图2所示.

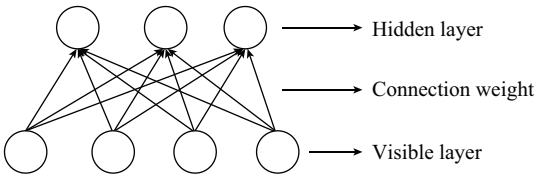

图 2　受限玻尔兹曼机的结构图

Fig. 2　Structure diagram of the RBM

由RBM层间有连接, 层内无连接的特殊结构可看出: 当给定可见单元的状态时, 各隐单元的激活状态之间条件独立. 其中第1层称为可视层, 用于数据特征的输入, 另一层是隐含层, 也就是特征提取层, W是可视层与隐含层之间的权重矩阵.

3.2 深度信念网络结构

深度信念网络又译为深层信念网络, 由Hinton及其合作者在2006年提出[11], 可以认为它是一种多隐层的人工神经网络, 其结构示意图如图3所示. 其结构由RBM堆叠构造而成, 其中的每一层RBM都可以单独作为聚类器, 可以用来对数据的概率分布进行建模, 也可以用来对数据进行分类. DBN的神经元可以分为显性神经元和隐性神经元, 显性神经元用于数据的输入, 隐性神经元用于数据特征的提取. 此外, DBN作为判别模型时, 深度信念网络的学习过程更像自编码器, 在经过受限玻尔兹曼机的逐层训练之后, 直接再用反向传播(Back propagation, BP)进行有监督的调优, 无须睡醒算法的参与.

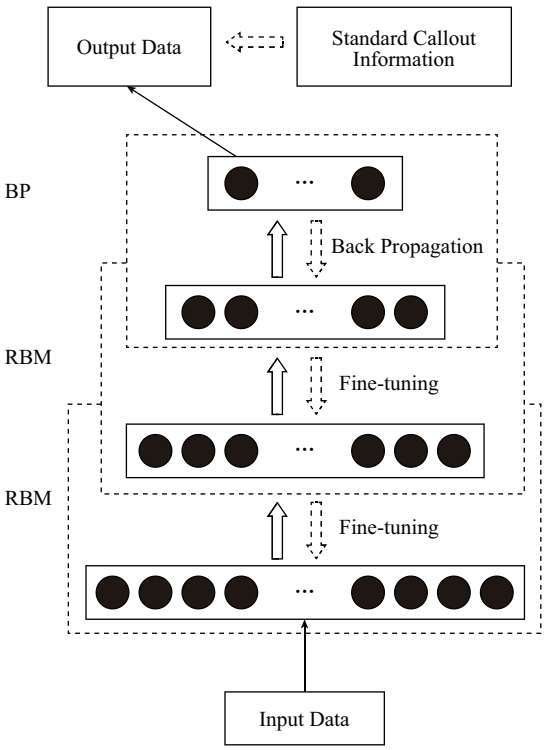

图 3 深度信念网络的结构图

Fig. 3 Structure diagram of the DBN

DBN得益于模型本身的构造其最主要的优势就是对数据特征的分层学习, 也就是对深度学习算法而言, 本身结构就具有良好的降维功能和性能, 使得能够从大数据中学习到合适且有效的特征. 因此, 在DBN结构的构造过程中, 虽然不需要像传统的分类算法一样进行显示特征的提取, 只是针对光谱数据, 也需要考虑相应的DBN模型中各层网络的选取和构造, 从而取得较好的特征学习能力和降维能力(在用传统方法构造光谱分类器时, 光谱特征的提取和选择是非常重要的一项工作. 可通过测量特征谱线的参量,

例如谱线的线心深度、等值宽度、特征谱线最大相对强度、特征谱线的特征波长、特征谱线的辐射强度度量等作为特征, 以降低光谱数据的维度).

3.3 深度信念网络的算法过程

(1) DBN的预训练过程: 分别对每一层RBM网络进行单独无监督训练, 使其数据的特征在不同空间的映射过程中, 都尽量保留光谱数据的特征信息.

(2) DBN的微调过程: 在DBN的结构中, 前面的每一层RBM网络都只能使得自身层内的权值对该层特征向量映射达到最优, 并不是对整个DBN的特征向量映射达到最优. 因此需要设置最后一层BP网络层, 将错误的信息自顶向下传播至每一层RBM层, 再全局微调整个DBN网络, 这样的训练过程使DBN克服了BP网络因随机初始化权值参数而容易陷入局部最优和时间复杂度高的问题.

3.4 深度信念网络模型的参数选择

DBN中RBM的层数越多对应的学习次数也越多, 获得的学习效果随之变好. 这里我们采用分类精确率和时间复杂度两个指标来对模型进行综合评估, 精确率的计算遵循以下计算方法:

$$A = \frac{1}{N}\sum_{i=1}^{N}\left\{f\left(x^i\right) = y^i\right\}, \tag{2}$$

其中, A表示样本数据得到的分类精确率, N为样本个数, y^i和x^i分别表示样本的标签类别与模型输出的类别.

此外, 由于我们采用的模型不需要设置学习控制参数, 比如步长和动量的设置, 只需要设置隐含层的节点个数. 为了确保结果的可靠性, 所涉及到的DBN模型中参数设置均保持一致: 输入的光谱数据维度均为3909, RBM层数为3层, 类别数为3类. 其中, 第1层RBM的可见元个数设置为500, 第2层RBM的隐含层节点个数设置为500, 第3层RBM的隐含层节点个数设置为2000, 即DBN网络节点数分别为3909-500-500-2000-3. 训练中迭代次数为200次, 学习率为0.1.

4 分类实验与结果分析

与我们方法不同的是Liu等[10]采用27种线指数特征和支持向量机对LAMOST光谱数据进行自动分类研究(线指数是一个用谱线名称命名的数值, 用这个数值来代表光谱中的某些物理特征. 它可以是光谱中某一段的积分星等, 可以是某条谱线的等值宽度(EW)或者半高全宽(CFWHM), 也可以是光谱中几个线指数的组合), 利用线指数来描述光谱特征, 在对高维数据进行降维时能够较完整地保留光谱信息. 在此基础上, 文章基于支持向量机进行分类研究, 结果显示对G型光谱能够被很好地分类, 但对于F和K光谱分类效果并不十分理想. 这可能有3个原因:

(1)晚-F型和早-G型光谱, 晚-G型和早-K型光谱的谱线特征很相似, 在SVM中很难区分, 见图4和图5;

(2)所采用的27种线指数特征对于高维光谱而言特征提取可能不具有代表性, 使得高维光谱数据的特征有所丢失;

(3)从SIMBAD数据库中采用的正确类是从多种文献以及用肉眼分类收集到的数据,

彼此之间很难校准.

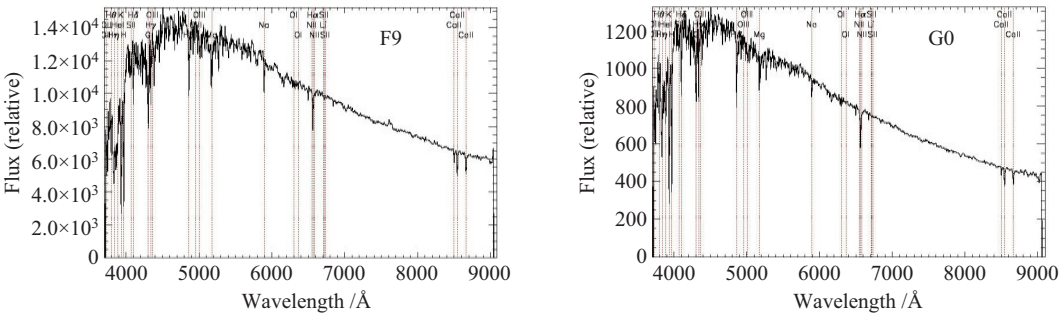

图 4 晚-F型与早-G型星谱线特征图

Fig. 4 Spectral lines feature of the late-F and early-G type

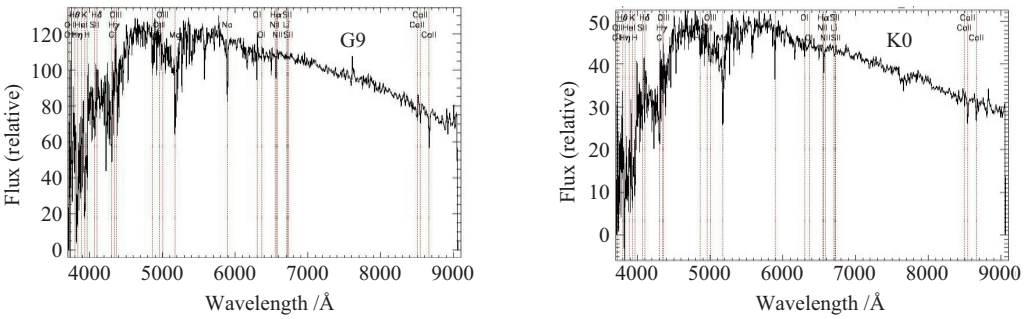

图 5 晚-G型与早-K型星谱线特征图

Fig. 5 Spectral lines feature of the late-G and early-K type

本实验选取的DBN模型最大的优势在于对光谱数据特征的分层学习本身就具备降维功能, 能够很好地提取光谱数据的显示特征, 从而更好地进行特征学习和分类实验. 基于以上样本和分类模型进行分类实验, 并将分类结果与Liu等[10]的结果进行对比分析, 结果见表1.

表 1 小样本的分类结果比较

Table 1 Comparison of classification results of small samples

Spectral types	Accuracy (Liu et al.[10])	Accuracy (DBN model)
F	0.72	0.8891
G	0.9091	0.9377
K	0.52	0.8379

考虑到以上实验样本数据较少, 结果可能不具有代表性, 而DBN模型的优势又在于样本越大分类精确度越高. 因此, 我们引入了更具有代表性的大样本, 样本总数为30000条光谱数据, 维度为3909, 分别为F、G、K型. 值得注意的是, 除对数据进行归一化处理以外, 未进行其他预处理过程, 且未限制光谱的信噪比值, 样本标签分别为1、2、3, 且在

实验中将样本分为训练集和测试集, 其中训练集27000条光谱, 测试集3000条光谱. 此外, Wang等[13]采用深度神经网络对同样F、G、K 3种类型恒星光谱的分类结果进行比较, 样本为30000条, 对光谱信噪比没有作限制. 文章采用深度神经网络分类模型, 节点设计为721-400-800-1200-2000-3, 即有4个隐含层的分类器模型. 可见在光谱数据保留721个特征的基础上基于伪逆学习算法对光谱进行分类实验, 分类精确率为0.819. 本实验同样采用深度学习模型, 但与其不同的是深度信念网络结构由受限玻尔兹曼机堆叠而成, 不需要对光谱进行降维, 应用该模型对高维光谱数据的特征分层学习能力, 尽可能保留有效特征以提升分类精确度. 基于以上模型进行分类实验, 并将所得结果同样在TOPCAT中与GAIA数据交叉, 得到的颜色-星等图如图6所示, 且将分类精确度与文献[11]进行比较, 结果见表2 (注: PILDNN是指基于伪逆学习算法的深度神经网络, PILDNN*是指将每条光谱作为输入向量时分为4个阶段并基于伪逆学习算法的深度神经网络).

表 2 大样本的分类结果比较

Table 2 Comparison of classification results of large samples

Classifier	Data dimension	Spend time/s	Accuracy
RBM	721	1481.0040	0.7572
PILDNN	721	226.9495	0.8190
PILDNN*	721	1103.4251	0.8232
DBN	3909	8611.0092	0.9303

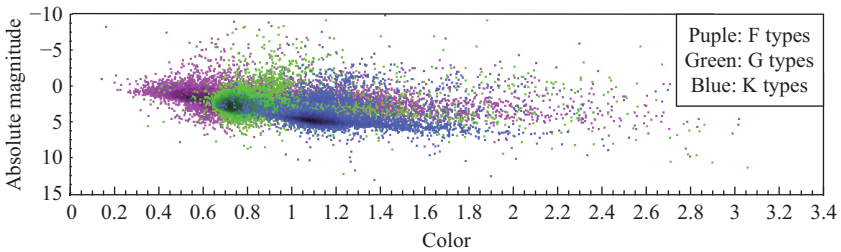

图 6 输出结果的颜色-星等图

Fig. 6 The color-magnitude diagram of output results

本文采用DBN对F、G和K型恒星光谱进行分类研究, 由以上结果分析可知: 当样本较小时, 3种类型的光谱在分类精确度上均有较大提升. 在大样本分类工作中, 图7、图8和图9分别是F、G和K型恒星光谱的输入输出颜色-星等图. 由此可知: 采用DBN模型对F、G和K 3种类型的光谱进行分类识别, 在精确率上均有很大提升, 尤其针对K型光谱而言, 参考文献中的精确度只达到0.52, 但就该模型而言, 采用分层学习特征的方法, 尽可能地保留了光谱数据的显著特征用于分类识别工作, 得到的结果更具说服力.

鉴于LAMOST巡天项目发布的大样本数据优势, 将DBN分类模型用于F、G、K 3种光谱数据中, 由以上分析可以看出:

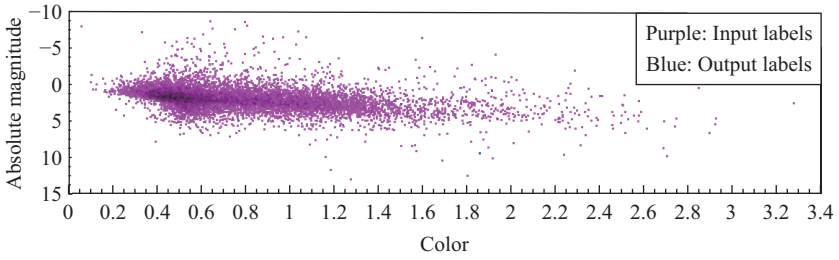

图 7 F型光谱输出类别的颜色-星等图

Fig. 7 The color-magnitude diagram of the F-type stars output labels

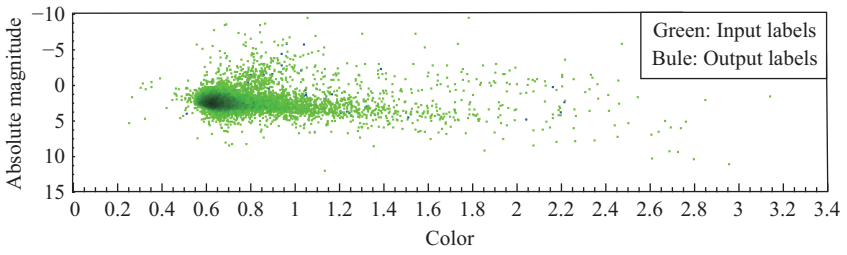

图 8 G型光谱输出类别的颜色-星等图

Fig. 8 The color-magnitude diagram of the G-type stars output labels

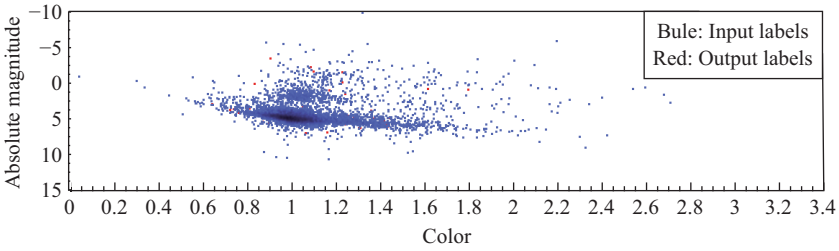

图 9 K型光谱输出类别的颜色-星等图

Fig. 9 The color-magnitude diagram of the K-type stars output labels

(1) DBN分类模型与其他算法相比较, 该模型通过受限玻尔兹曼机分层学习、训练各个参数的权值, 并根据目标函数值经误差反馈对参数数值进行微调, 使得对于天体光谱的总体分类精确率有明显提升;

(2) DBN模型充分体现了样本越大, 分类准确率越高的优势. 由于DBN结构的特殊性, 样本越大, 在训练过程中越能分层学习到更多有效的光谱特征信息, 进而提升分类准确率;

(3) DBN模型具有较强的学习能力, 不需要对数据进行降维处理, 且精确率优于降维处理后的效果. 常见的分类算法(SVM、BP神经网络和终端学习机(ELM)等算法)往往存在陷入局部最优或者过度学习的问题, 为了避免维数灾难通常需要对数据做降维处理, 因此得到的结果并不能很好地反映数据的总体特征, 最终导致分类准确率的下降. 而DBN模型可以从高维的原始数据中提取差别较大的低维特征, 不需要对数据进行降

维就可直接开始训练分类模型, 不仅能够更全面地考虑到光谱信息量, 而且能够较为准确地对光谱数据进行分类识别.

5 总结与展望

文章针对原有方法分类可信度较低的F、G、K 3种类型的恒星光谱, 采用DBN进行大样本分类实验, 结果表明通过分层学习提取光谱数据特征的方法具有很好的鲁棒性, 且分类效果优于其他分类模型. 深度学习方法虽然在大样本数据分类识别时具有较大优势, 但是该方法计算量巨大, 对计算资源具有较高要求, 因此, 还需要优化算法以解决计算复杂度高的问题. 在接下来的工作中, 我们会继续选取分类精度低或光谱型未知的光谱作为分类搜寻的候选体, 采用DBN模型输出分层学习得到的特征, 与专家选择的特征进行比较, 进而分析模型的有效性. 此外, 本文所采用的有监督学习都是基于强标签标记的数据, 后续工作中我们将搜寻弱标签样本, 采用该模型进行研究. 并在此基础上, 我们也将继续采用聚类分析和离群点分析, 对特殊天体进行观测证认, 进一步完备各型巨星样本. 除此之外, 还计划采用数据挖掘技术中的关联规则挖掘方法, 尝试发现各类已知和未知的天体光谱之间的联系, 进一步揭示尚未被发现的天体物理规律, 研究成果可以为银河系结构和动力学研究提供更好的支持.

参 考 文 献

[1] York D G, Adelmanet J, Anderson J E, et al. AJ, 2000, 120: 1579
[2] Perryman M A C, De Boer K S, Gilmore G, et al. A&A, 2001, 369: 339
[3] Cui X Q, Zhao Y H, Chu Y Q, et al. RAA, 2012, 12: 1197
[4] 吴永东, 马颂德. 中国图象图形学报, 1997, 2: 3
[5] 覃冬梅, 胡占义, 赵永恒. 光谱学与光谱分析, 2003, 23: 182
[6] 覃冬梅, 胡占义, 赵永恒. 光谱学与光谱分析, 2004, 24: 507
[7] Shi F, Liu Y Y, Sun L G, et al. MNRAS, 2015, 453: 122
[8] 赵瑞珍, 王飞, 罗阿里, 等. 光谱学与光谱分析, 2009, 29: 2010
[9] Luo A, Zhao Y H, Zhao G, et al. RAA, 2015, 15: 1095
[10] Liu C, Cui W Y, Zhang B, et al. RAA, 2015, 15: 1137
[11] Hinton G E, Salakhutdinov R R. Science, 2006, 313: 504
[12] Ackley D H, Hinton G E, Sejnowski T J. Cognitive Science, 1985, 9: 147
[13] Wang K, Guo P, Luo A L. MNRAS, 2016, 465: 4311
[14] 潘儒扬, 李乡儒. 天文学报, 2016, 57: 379
[15] Pan R Y, Li X R. ChA&A, 2017, 41: 318

Classification of LAMOST Spectra Based on Deep Learning

XU Ting-ting[1] MA Chen-ye[2] ZHANG Jing-min[1] ZHOU Wei-hong[1,3]

(1 School of Mathematics and Computer Sciences, Yunnan Minzu University, Kunming 650500)
(2 School of Economics and Business Administration, Yunnan Vocational and Technical College of Agriculture, Kunming 650031)
(3 Key Laboratory for the Structure and Evolution of Celestial Objects, Chinese Academy of Sciences, Kunming 650011)

Abstract The classification of the normal stars plays important roles not only in the understanding of the stellar physics, but also in the study of the overall structure and evolution of the Milky Way. However, the problems of low classification accuracy or unknown spectral types still exist in the related researches. In this paper, we present a new automated feature extraction method for spectra with its application in spectral classification of F, G, and K type stars. The basic idea of our approach is to train a deep belief network to hierarchical learning features of spectra data, and then we use the back propagation algorithm to fine-tuning the whole model. We select 31667 spectra that include three type stars of F, G, and K from LAMOST (the Large Sky Area Multi-Object Fiber Spectroscopic Telescope) Data Release 5, and cross with GAIA (the Global Astrometric Interferometer for Astrophysics) data to get the color-magnitude diagram in the TOPCAT. We evaluate the performance of the proposed scheme and the results demonstrate that the method of deep belief network is superior in the comprehensive performance.

Key words stars: fundamental parameters, methods: data analysis, methods: statistical, techniques: spectral analysis