doi: 10.15940/j.cnki.0001-5245.2021.01.008

基于胶囊网络的恒星光谱分类研究*

杜利婷[1†] 洪丽华[2] 杨锦涛[1] 许婷婷[3] 张静敏[1] 艾霖嫔[1]
周卫红[1,4‡]

(1 云南民族大学数学与计算机科学学院 昆明 650500)
(2 厦门软件职业技术学院软件工程系 厦门 361000)
(3 广州大学天体物理中心 广州 510006)
(4 中国科学院天体结构与演化重点实验室 昆明 650011)

摘要 大型巡天项目的快速发展, 产生大量的恒星光谱数据, 也使得实现恒星光谱数据的自动分类成为一项具有挑战性的工作. 提出一种新的基于胶囊网络的恒星光谱分类方法, 首先利用1维卷积网络和短时傅里叶变换将来源于LAMOST (Large Sky Area Multi-Object Fiber Spectroscopy Telescope) Data Release 5 (DR5)的F5、G5、K5型1维恒星光谱转化成2维傅里叶谱图像, 再通过胶囊网络对2维谱图像进行自动分类. 由于胶囊网络具有保留图像中实体之间的分层位姿关系和无需池化层的优点, 实验结果表明: 胶囊网络具有较好的分类性能, 对于F5、G5、K5型恒星光谱的分类, 准确率优于其他分类方法.

关键词 恒星: 基本参数, 方法: 数据分析, 技术: 光谱分析
中图分类号: P144; **文献标识码**: A

1 引言

随着巡天项目的快速发展, 如斯隆数字巡天(Sloan Digital Sky Survey, SDSS)[1], LAMOST (Large Sky Area Multi-Object Fiber Spectroscopy Telescope)光谱巡天[2], 每年产生海量天文数据, 天文大数据为天文学家研究银河系及一般星系的形成与演化提供了有力的基础性数据, 如何对海量的恒星光谱数据进行准确识别分类成为一大难题.

2017年6月, LAMOST圆满完成了为期5 yr的第1期低分辨率($R = 1800$, R为光谱分辨率)光谱巡天任务, 2017年12月31日, LAMOST Data Release 5 (DR5)数据集正式发布, 共包括4154个观测天区, 发布了901万条光谱, 其中高质量光谱数(信噪比大于10)达到了777万条, 远超过全世界光谱巡天项目获取的光谱数总和; 同时, DR5发布的数据中还有一个包括636万组恒星光谱参数的星表, 成为目前全世界最大的恒星参数星表.

2020-06-26收到原稿, 2020-08-21收到修改稿
*国家自然科学基金项目(61561053), 云南民族大学数学与计算机科学学院研究生科研项目(SJXY 2020-103、SJXY2020-101)资助
†2310514572@qq.com
‡ynzwh@163.com

2017年9月至2018年6月, LAMOST处于中分辨率($R = 7500$)测试观测期, 即低分辨率光谱巡天和中分辨率测试观测交替进行的观测模式. 2019年3月发布的包含先导巡天及前6 yr正式巡天的LAMOST Data Release 6 (DR6)数据集包括常规低分辨率光谱数据和中分辨率测试光谱数据, 共4902个观测天区、1125万条光谱, 其中低分辨率光谱数据总数991万条, 中分辨率非时域光谱数据50万条, 中分辨率时域光谱数据84万条; 信噪比大于10的高质量光谱数量达到了937万条, 至此, 巡天7 yr的LAMOST成为世界上第1个获取光谱数突破千万量级的光谱巡天项目, 标志着LAMOST光谱发布正式进入千万量级时代.

深度学习是机器学习中非常接近人工智能的领域, 其动机在于建立模拟人脑进行分析学习的神经网络, 优点是面对大样本数据学习能力强, 随着深度学习的不断发展, 其应用面也不断扩大, 也包括对恒星光谱的分类.

1943年, McCulloch等[3]提出并给出了人工神经网络的概念及人工神经元的数学模型, 从而开创了神经网络研究的时代. 到1958年, Rosenblatt首次提出了可以模仿人类感知能力的机器, 并称之为感知机[4], 感知机是有单层计算单元的神经网络, 由线性元件及阈值元件组成, 最大作用就是对输入的样本分类, 故它可以作为分类器, 是整个神经网络的基础. 1982年, 误差反向传播(Back Propagation, BP)算法解决了感知机隐含层的权值问题, 它的基本思想: 学习过程由信号的正向传播与误差的反向传播两个过程组成, Bailer-Jones等[5–6]采用主成分分析(Principal Component Analysis, PCA)降维并结合BP神经网络对恒星光谱进行了系统分类, 识别结果较好; Qin等[7]采用PCA方法对恒星光谱数据进行分类, 实验结果表明, 该方法可以达到与摩根-基南(Morgan-Keenan, MK)系统[8]分类准则相当的性能, 可作为天文领域的一个基准. 基于树的机器学习算法是一类有监督的机器学习算法, 具有模型结构相对简单、运算量相对较小, 同时准确率相对较高等优点. 其中, Chen等[9]提出的极端梯度提升(eXtreme Gradient Boosting, XGBoost)算法是一种迭代型树类算法, 其更容易实现并行处理、运算处理速度更快、比传统决策树算法准确性更高, 因而备受瞩目成为一种流行的机器学习算法. Zhang等[10]对源于LAMOST Data Release 4 (DR4)的B、A、F、M型恒星光谱进行分类. 首先对光谱数据计算谱线指数从而使其得到降维处理, 过滤冗余信息, 然后通过XGBoost算法得到分类器模型再对降维后的光谱数据进行分类. 通过实验可以发现, 在固定参数下, XGBoost所得的模型有一定的自适应性, 总体准确率可达88.5%; 潘景昌等[11]提出了基于Lick线指数的贝叶斯光谱分类方法, 首先基于Hadoop平台计算各类光谱的Lick线指数作为特征向量, 然后利用贝叶斯分类算法对F、G、K三类恒星光谱进行分类. 2006年, Hinton提出了针对深层网络训练中梯度消失问题的解决方案: 无监督预训练对权值进行初始化和有监督训练微调, 其主要思想是先通过自学习的方法学习到训练数据的结构(自动编码器(Auto Encoder, AE)), 然后在该结构上进行有监督训练微调. 许婷婷等[12–13]利用深度信念网络(Deep Belief Network, DBN)对F、G、K型恒星光谱进行分类研究, 分类准确率达到93.03%. 2012年, Hinton为了证明深度学习的潜力, 首次参加ImageNet图像识别比赛, 通过他构建的AlexNet卷积神经网络(Convolutional Neural Network, CNN)一举夺得冠军, 其分类性能完全碾压了获得第2名的支持向量机(Support Vector Machine, SVM)方法. Shi等[14]针对恒星光谱自动分类问题, 提出了一

种基于CNN的K和F型恒星光谱自动分类方法, 并与SVM和BP算法进行对比, 对比实验结果表明, CNN算法明显优于SVM和BP算法.

文中结构如下: 第2节介绍胶囊网络; 第3节介绍实验步骤的设计; 第4节为实验结果的分析; 最后为结语.

2 胶囊网络

深度学习中的CNN, 其权值共享网络结构可以显著降低模型复杂度, 减少权值数量; 图片可以直接作为网络的输入, 自动提取特征, 在图像处理中具有很大的优势, 但是也有许多不足之处: 首先, CNN无法从新的视角去理解对象, 当图片角度发生变化时CNN就无法识别, CNN如果通过数据增强方式构成新的视角(数据增强方式包括将图片进行翻转、旋转等操作)就会造成数据量非常大, 从而降低训练速度; 其次, 难以精确识别空间关系, 如果将图片中的某个要素移动位置后, CNN可能就无法识别; 最后, CNN中的最大池化虽然可以通过减少网络空间大小来获得计算量更小的优势, 但是同时也造成了大量的信息丢失, 最终导致分类准确率下降. 因此, 针对CNN以上不足, 提出了胶囊网络.

胶囊网络的一般结构包括输入层、卷积层、主胶囊(PrimaryCaps)层、数字胶囊(DigitCaps)层、全连接层、输出层. 首先, 在输入层放入数据集, 进行相关的预处理; 其次, 在卷积层处用卷积核提取特征得到特征图; 在主胶囊处, 一般会拥有多个胶囊, 每个胶囊都是包含多个神经元的载体, 每个神经元表示图像中出现的特定实体的各种属性, 比如位置、大小、方向等, 即将胶囊类比于向量, 长度代表特定实体在图像某个位置存在的概率, 方向代表特定实体的一些参数, 比如位置、大小、转角等, 卷积层处得到的特征图进入主胶囊层后会在每个胶囊中用相同的卷积核再次对特征图进行特征提取, 形成新的特征图并以向量形式表示; 然后进入数字胶囊层, 主胶囊层与数字胶囊层是全连接的, 但是以向量形式相连, 使用动态路由算法完成权重更新和实现从主胶囊层到数字胶囊层的输出, 通过计算向量模的大小来衡量某个实体出现的概率, 模值越大概率越大; 在全连接层处进行重构; 最后输出相关结果. 卷积神经网络和胶囊网络的一般结构如图1所示, 在卷积神经网络中的convolution 1表示卷积处理; 胶囊网络中的convolution 1表示卷积处理, ReLu表示激活函数.

在胶囊网络中, 低级别特征(主胶囊层)通过改变权重c_{ij}将其输出向量发送到高级别特征(数字胶囊层)处, 其中i表示主胶囊层处的胶囊序号, j表示数字胶囊层层处的胶囊序号, c_{ij}即表示胶囊i激活胶囊j的权重分布, 在这个过程中, 权重的更新是由迭代动态路由算法来实现, 动态路由算法的步骤如表1, 其中涉及到的公式如下:

$$\boldsymbol{u}_{j|i} = \boldsymbol{w}_{ij} \cdot \boldsymbol{u}_i \,, \tag{1}$$

$$c_{ij} = \frac{\exp(b_{ij})}{\sum_k \exp(b_{ik})} \,, \tag{2}$$

$$\boldsymbol{s}_j = \sum_i c_{ij} \boldsymbol{u}_{j|i} \,, \tag{3}$$

$$\boldsymbol{v}_j = \frac{\|\mathbf{s}_j\|^2}{1+\|\mathbf{s}_j\|^2} \cdot \frac{\mathbf{s}_j}{\|\mathbf{s}_j\|} \,, \tag{4}$$

$$b_{ij} = b_{ij} + \boldsymbol{u}_{j|i} \cdot \boldsymbol{v}_j \,. \tag{5}$$

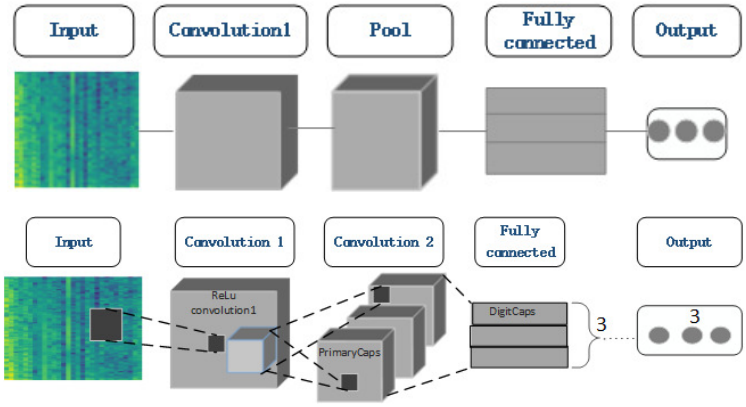

图 1 卷积神经网络和胶囊网络的结构对比

Fig. 1 Structure comparison of convolutional neural network and capsule networks

表 1 动态路由算法

Table 1 Dynamic routing algorithm

Procedure routing algorithm:

1: procedure ROUTING($\boldsymbol{u}_{j|i}$, r, l)

2: for all capsule i in layer l and capsule j in layer $(l+1)$: $b_{ij} \leftarrow 0$

3: for r iterations do

4: for all capsule i in layer l: $c_{ij} \leftarrow \mathrm{softmax}(b_{ij})$ ▷ softmax computes Eq. (3)

5: for all capsule i in layer $(l+1)$: $\boldsymbol{s}_j \leftarrow \sum_i c_{ij}\boldsymbol{u}_{j|i}$

6: for all capsule i in layer $(l+1)$: $\boldsymbol{v}_j \leftarrow \mathrm{squash}\ (\boldsymbol{s}_j)$ ▷ squash computes Eq. (1)

7: for all capsule i in layer l and capsule j in layer $(l+1)$: $b_{ij} \leftarrow b_{ij} + \boldsymbol{u}_{j|i} \cdot \boldsymbol{v}_j$

Return $\boldsymbol{v}_j$

在(1)式中, $\boldsymbol{u}_i$表示低层特征, 即胶囊i的输入, $\boldsymbol{w}_{ij}$表示低层特征与高层特征的空间关系, 通过反向传播进行学习, $\boldsymbol{u}_{j|i}$表示由低层特征推出的高层特征; 在(2)式的归一化函数中, c_{ij}表示权重, 它的特点有4个, 分别为均为非负标量, 所有权重之和为1, 权重个数为胶囊数量, 权重由迭代路由算法确定; k表示胶囊i内的神经元个数; 在(3)式中, 进行一个加权求和; 在(4)式中, 利用压缩函数对$\boldsymbol{s}_j$进行压缩, $\boldsymbol{v}_j$是胶囊j的输出向量, $\boldsymbol{s}_j$是它的全部输入; 在(5)式中, 利用相似函数对b_{ij}进行更新.

动态路由算法逻辑解释:

第1行: 迭代次数r (文中为3), 在l层输入$\boldsymbol{u}_i$的输出$\boldsymbol{u}_{j|i}$; 第2行: 初始化所有b_{ij}为0, b_{ij}表示从胶囊i应该耦合到胶囊j的对数先验概率, 取决于胶囊i与j的类型与位置, 与当前输入图像无关; 第3行: 执行4到7行3次; 第4行: 对l层的低层特征, 将b_{ij}用softmax转化成权重c_{ij}, softmax函数产出是非负数且总和为1, 这使得c_{ij}是一组概率变量; 第5行: 对l+1层的高层特征, 加权求和得到$\boldsymbol{s}_j$; 第6行: 对l+1层的高层特征, 利用squash压缩$\boldsymbol{s}_j$得到$\boldsymbol{v}_j$, squash函数(压缩函数)确保向量$\boldsymbol{s}_j$和$\boldsymbol{v}_j$的方向相同, 且$\boldsymbol{v}_j$长度不超过1; 第7行: 根据

$\boldsymbol{u}_{j|i}$和$\boldsymbol{v}_j$的点积来更新b_{ij}, 两者相似, 点积就越大, b_{ij}就越大, 低层特征连接高层特征的可能性就变大; 反之, 两者相异, 点积就越小, b_{ij}就越小, 低层特征连接高层特征的可能性就变小.

3 实验设计

3.1 利用1维卷积网络和短时傅里叶变换生成2维傅里叶谱图像

3.1.1 1维卷积网络

CNN属于人工神经网络的一种, 神经网络的基本组成包括输入层、隐藏层、输出层. CNN的特点在于隐藏层分为卷积层和池化层, 卷积过程可以用来消除噪声、增强特征, 池化层减少参数进而降低网络的复杂度. 对于2维CNN结构, 输入的是2维矩阵, 这种结构经常用在图像识别上, 因为图像有宽和高2个维度, 但是对于1维的恒星光谱数据, 2维CNN无法使用, 所以对恒星光谱数据进行只包括输入层、卷积层、输出层的1维卷积过程处理, 其卷积核只在一个方向上进行滑窗操作, 即在宽方向或高方向上进行加权求和, 且进行填充以保持光谱数据原有维度不变, 从而相应特征也保留, 继而减少在短时傅里叶变换采样过程中造成的信息损失.

在1维卷积结构处, 首先在输入层输入恒星光谱数据F5、G5、K5各1000条; 而在卷积层处, 我们设置了数量为1, 尺寸为3×1, 权值为1、1、1, 步长为1的卷积核, 并进行填充, 起到平滑的效果. 具体计算过程如图2; 最后在输出层输出经过卷积层后的结果.

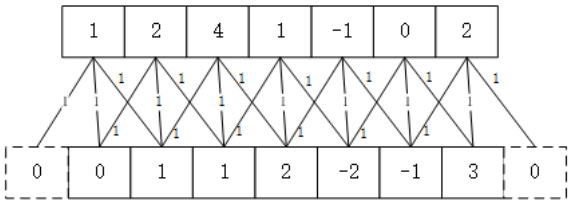

图 2 1维卷积的运算

Fig. 2 Operation of one-dimensional convolution

3.1.2 短时傅里叶变换

傅里叶变换(Fourier Transformation, FT)通过将信号分解成正余弦函数, 将时域信号转化为频域信号, 并提取到在空域上不易提取到的特征. 但是不足之处是其只适用于平稳信号, 在频域图上不能获得对应频率的时间信息, 为此提出短时傅里叶变换(Short Time Fourier Transform, STFT), 设置窗格, 认为窗格内的信号是平稳信号, 对窗格内的信号分段进行傅里叶分析, 得到一系列频域信号的变化结果, 将这些结果排开便得到一个2维的表象, 进而再用相关工具去分析, 得到某个时段上的频率特征, 实现时频局部化. STFT的公式如下:

$$\mathrm{STFT}_Z(t,f)=\int_{-\infty}^{+\infty}[Z(u)g^*(u-t)]\exp(-q2\pi fu)\mathrm{d}u\,,$$

其中, u为某个时间段, $Z(u)$为源信号, $g^*(u-t)$是一个中心为t的窗函数, f是信号函数中的基频率, q为虚数单位.

实验利用STFT的Specgram函数将1维恒星光谱数据转换成2维傅里叶谱图像, 形成新的特征分布且保留了更多特征, 有利于后续的分类研究. Specgram函数如下:

$$[S, F, T, P] = \text{Specgram}(\boldsymbol{a}, \text{window}, \text{noverlap}, \text{nfft}, \text{fs}) ,$$

其中, $\boldsymbol{a}$是输入信号的向量; window是窗函数, 默认为nsc, nsc表示海明窗的长度; noverlap是每相邻两个窗口的重叠率; nfft是每个窗口的快速傅里叶变换采样点数; fs是采样频率(本文为5); S是信号a的短时傅里叶变换; F是在输入变量中使用F频率变量; T是频谱图计算的时刻点; P是能量谱密度.

实验中使用Python对恒星光谱数据依次进行1维卷积处理和STFT. 一条原始F5型恒星光谱数据以及该条恒星光谱数据经过STFT后生成的2维傅里叶谱图像如图3所示.

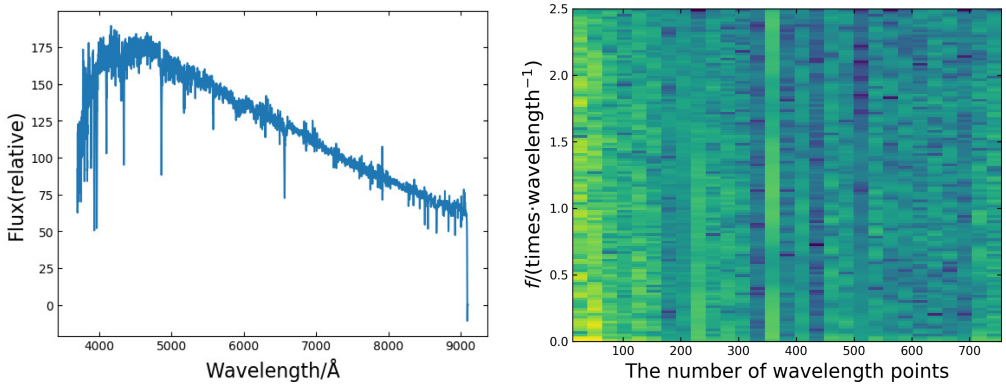

图 3 原始恒星光谱数据和其经过STFT后生成的2维傅里叶谱图像

Fig. 3 Original star spectral data and its two-dimensional Fourier spectrum image generated after STFT

3.2 基于胶囊网络对2维谱图像的分类

将F5、G5、K5型恒星光谱数据对应的2维傅里叶谱图像作为胶囊网络的输入, 依次经过卷积层、主胶囊层、数字胶囊层、全连接层, 最后得到分类结果. 在本实验中, 卷积层处设置256个步长为1的9 × 9的卷积核; 在主胶囊层处设置8个胶囊, 32个步长为2的9 × 9 × 256的卷积核; 在数字胶囊处会输出16 × 3的矩阵, 3是因为有F5、G5、K5 3个类别, 每个元素是1 × 16的向量; 接下来进入3个全连接层, 在最后一个全连接层处, 得到重整后重建1920 × 1440的解码图像, 损失函数为重建图像和输入图像之间的欧氏距离, 最后得到分类结果.

4 结果分析

4.1 实验环境

在AMD A8的处理器下进行, 实验平台为Python 3.0.

4.2 实验数据

文中实验数据来源于LAMOST DR5, 从中随机选取各1000条共3000条F5、G5、K5型恒星光谱数据, 信噪比> 20, 每条光谱数据的波长范围是3700–9100 Å.

4.3 实验结果与分析

在文献[15–16]中, 张静敏等人通过短时傅里叶变换(STFT)将1维恒星光谱数据转换成新的特征谱图像, 再利用Inception v3模型对2维傅里叶谱图像进行分类实验, 实验数据来源于LAMOST DR5中的F型、G型、K型各10000条, 最终的分类准确率为92.9%; 在之前的实验中, 提出在利用短时傅里叶变换(STFT)之前, 先将来自LAMOST DR5的各1000条F5、G5、K5型恒星光谱数据做1维卷积处理, 以减少在短时傅里叶变换的采样过程中造成的特征损失, 然后再利用STFT将恒星光谱数据转化为2维傅里叶谱图像, 最后利用Inception v3模型对2维傅里叶谱图像进行分类, 分类准确率为99%.

基于文献[15–16]与之前实验中的分类器都是属于CNN中的经典模型Inception v3, 但CNN在图像处理中存在许多不足, 而胶囊网络正好可以解决这些不足, 因此将分类器替换成胶囊网络, 结果如表2所示.

表 2 结果比对
Table 2 The comparison of results

Type of star	The number of sample	1D convolution	Classifier	Accuracy/%
F5, G5, K5	3000	Yes	Inception v3	99
F5, G5, K5	3000	Yes	Capsule Network	99.67

通过对比我们可以发现, 胶囊网络对恒星光谱数据的分类准确率高于Inception v3, 提高了恒星光谱的分类准确率, 验证了胶囊网络相较于CNN具有的优势, 能对恒星光谱数据进行有效分类.

相较于经典的CNN, 胶囊网络实现了3个不同: 首先是在对象部件间的分层位姿关系建模上的区别. 在CNN中, 通过训练神经元来检测不同的实体, 即使是同一实体的不同角度, 这样使得卷积核的个数和层数越来越多, 而在胶囊网络中, 其通过一个胶囊就能够识别同一类实体, 胶囊输出向量的长度代表目标存在的概率估计, 向量的方向代表实体的属性; 其次是设置动态路由算法有所不同. 其采用新型非线性向量激活函数squash来实现权重更新和从主胶囊层到数字胶囊层的输出; 最后是使用胶囊网络取代了CNN中的最大池化, 在减少了数据空间的同时, 又尽量保障了重要信息的传递. 这两个创新之处加强了对图像特征的提取并保留了更多的信息, 而实验结果也证实了相对于CNN, 胶囊网络有效提高了恒星光谱的分类准确率.

5 结语

本文在对已有的恒星光谱分类方法进行深入研究和总结的基础上, 提出了基于胶囊网络的恒星光谱分类方法, 对LAMOST DR5的F5、G5、K5型恒星光谱数据进行分类, 分类准确率达到99.67%, 从而验证胶囊网络能对恒星光谱数据进行有效分类. 但是, 本实验也存在不足, 比如胶囊网络目前的参数是否是最优还有待进一步验证; 其次是胶囊网络对计算机配置的要求较高, 计算量大导致耗费时间较长等不足. 在接下来的工作中, 将进一步对胶囊网络进行探究, 以提高分类准确率, 进而完备恒星光谱的数据库, 为研究恒星以及银河系的形成与演化提供支持.

参考文献

[1] York D G, Adelman J, Anderson J E, et al. AJ, 2000, 120: 1579
[2] Su D Q, Wang Y N. AcApS, 1997, 17: 315
[3] McCulloch W S, Pitts W. The Bulletin of Mathematical Biophysics, 1943, 5: 115
[4] Rosenblatt F. Psychological Review, 1958, 65: 386
[5] Bailer-Jones C A L, Irwin M, Gilmore G, et al. MNRAS, 1997, 292: 157
[6] Bailer-Jones C A L, Irwin M, Von Hippel T. MNRAS, 1998, 298: 361
[7] 覃冬梅, 胡占义, 赵永恒. 光谱学与光谱分析, 2003, 23: 182
[8] Morgan W W, Keenan P C. ARA&A, 1973, 11: 29
[9] Chen T Q, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016: 785
[10] 张枭, 罗阿里. 光谱学与光谱分析, 2019, 39: 3292
[11] 潘景昌, 王杰, 姜斌, 等. 光谱学与光谱分析, 2016, 36: 2651
[12] 许婷婷, 马晨晔, 张静敏, 等. 天文学报, 2019, 60: 13
[13] Xu T T, Ma C Y, Zhang J M, et al. ChA&A, 2019, 43: 353
[14] 石超君, 邱波, 周亚同, 等. 光谱学与光谱分析, 2018, 39: 1312
[15] 张静敏, 马晨晔, 王璐, 等. 天文学报, 2020, 61: 93
[16] Zhang J M, Ma C Y, Wang L, et al. ChA&A, 2020, 44: 334

Stellar Spectral Classification Based on Capsule Network

DU Li-ting[1] HONG Li-hua[2] YANG Jin-tao[1] XU Ting-ting[3]
ZHANG Jing-min[1] AI Lin-pin[1] ZHOU Wei-hong[1,4]

(1 School of Mathematics and Computer Science, Yunnan Minzu University, Kunming 650500)
(2 School of Software Engineering, Xiamen Institute of Software Technology, Xiamen 361000)
(3 Center for Astrophysics, Guangzhou University, Guangzhou 510006)
(4 Key Laboratory of the Structure and Evolution of Celestial Objects, Chinese Academy of Sciences, Kunming 650011)

Abstract The rapid development of large scale sky survey project has produced a large amount of stellar spectral data, which makes the automatic classification of stellar spectral data a challenging task. We select F5, G5, and K5 types of spectra data, which are derived from LAMOST Data Release 5 (DR5), one dimensional convolution network and short-time Fourier transform (STFT) are used to transform the one-dimensional spectra data into two-dimensional Fourier spectrum images. And then the two-dimensional Fourier spectrum images are classified automatically by the capsule network. Because the capsule network preserves the hierarchical pose relationships between the entities in the image and the advantages of removing the pooled layers, the experimental results show that for the classification accuracy of F5, G5, and K5 stellar spectra, capsule Network has better classification performance and is superior to other classification methods.

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