一种基于卷积神经网络的恒星光谱快速分类法

王楠楠[1]，邱　波[1*]，马　杰[1*]，石超君[1]，宋　涛[1]，郭　平[2*]

1. 河北工业大学电子信息工程学院，天津　300401
2. 北京师范大学系统科学学院，北京　100875

摘　要　恒星光谱数据的分类是天体光谱自动识别的最基本任务之一，光谱分类的研究能够为恒星的演化提供线索。随着科技的发展，天文数据也向大数据时代迈进，需要处理的恒星光谱数量越来越多，如何对其进行自动而精准地分类成为了天文学家要解决的难题之一。当前恒星光谱自动分类问题的解决方法相对较少，为此本文使用了一种基于卷积神经网络的方法对恒星光谱MK系统进行分类。该网络由数据输入层、四个卷积层、四个池化层、全连接层、输出层构成，与传统网络相比具有局部感知、参数共享等优点实验。在Python3.5的环境下编程，利用Tensorflow构建了一个简单高效的具有四个卷积层的卷积神经网络，并将Dropout作用于全连接层之后以防止过度拟合。Dropout的基本思想：当网络模型进行训练时，把一些神经网络节点按一定的比例丢弃，使其暂时不发挥作用。Dropout可以理解成是一种十分高效的神经网络模型平均方法，由于它不依赖于某些局部特征所以能够让网络模型更加鲁棒。实验中使用的一维恒星光谱图是取自LAMOST DR3数据库，首先进行预处理截取光谱3 600～7 300 Å的部分，均匀采样后使用min-max标准化法对其进行初始化。实验包括两部分：第一部分为依据恒星光谱MK系统对光谱进行分类，每一类的训练样本包含1 000条光谱数据，测试样本为400条光谱数据，首先通过训练样本对CNN网络进行训练，进行3 000次的迭代，用训练后的网络将测试样本进行分类以验证网络的准确性；第二部分为相邻两类的恒星光谱的分类，其中O型星数据集样本为250条光谱，其余类别恒星样本数据集均为4 000条光谱，将数据5等分，每次选取当中的一份当作测试集，其余部分当作训练集，采用5折交叉验证法求得模型准确率，用BP神经网络进行对比实验。选择对网络模型进行评估的指标包括精确率P、召回率R、F-score、准确率A。实验结果显示CNN在对六类恒星光谱进行分类时其准确率都在95%以上，在对相邻类别的恒星进行分类时，由于O型星样本量较少，所以得到的分类结果不太理想，对其余类别的恒星分类准确率都高于98%，以上结果都证明了CNN算法能够很好地解决恒星光谱的分类问题。

关键词　恒星光谱数据；自动分类；CNN；5折交叉验证

中图分类号：P157.2　　**文献标识码**：A　　**DOI**：10.3964/j.issn.1000-0593(2019)10-3297-05

引　言

随着多个巡天项目的进行，我们可以获得的光谱数据日益增多。如何从光谱中自动而准确地提取天文信息是当前天文学家面临的一个重要挑战[1]。LAMOST创新性地采用了主动光学技术，设计新颖，是目前天文望远镜中光谱获取率最高的。目前广泛使用的恒星分类方法称为Morgan-Kenan系统(MK系统)，属于二元分类系统[2]。MK系统分类依据的物理参量是温度和光度，恒星光谱可以按照温度递减分为O，B，A，F，G，K，M共7种。

近些年来国内外研究人员在光谱分类方面做了许多尝试，并且已经开发出了多种自动分类方法。Schierscher和Paunzen提出将人工神经网络(ANN)用于天文学中，作为光谱分类工具[3]，但是ANN的计算复杂度取决于输入空间的维数和隐藏层的大小，这意味着训练一个优秀的人工神经网络在计算上相当复杂。主成分分析(PCA)作为一种常用的降维方法也被天文学家用在光谱分类领域，Singh等应用PCA

收稿日期：2018-09-05，**修订日期**：2019-01-19

基金项目：国家自然科学基金委员会-中国科学院天文联合基金项目(U1531242)，河北省科技支撑计划(15212105D)，天津市企业科技特派员项目(18JCTPJC54300)资助

作者简介：王楠楠，1992年生，河北工业大学电子信息工程学院硕士研究生　　e-mail：814588655@qq.com

*通讯联系人　　e-mail：qiubo@hebut.edu.cn；jma@hebut.edu.cn；pguo@bnu.edu.cn

来降低光谱的维数，并应用多层 BP 网络(MBPN)实现分类过程的自动化[4]。由于 PCA 是一种线性降维方法，在对光谱降维时存在缺陷，所以 Daniel 等人提出将局部线性嵌入(LLE)应用到恒星光谱分类中[5]，然后将 LLE 的性能与 PCA 在光谱分类中的性能进行了比较，发现 LLE 优于 PCA。Liu 等将支持向量机(SVM)的分类算法应用到 LAMOST 恒星光谱分类的研究中，证实了使用 SVM 对光谱分类具有可行性[6]。实验结果表明该算法对 A，F，G 和 M 型恒星可以准确地分类，但在对 B 型或 K 型恒星进行分类时错误率接近 50%。薛建桥等提出了一种基于自组织特征映射(SOFM)进行光谱分类的研究方法，取得了与哈佛分类系统一致的分类结果[7]。

近些年来，深度学习在图片识别与分类、语音识别和医疗领域等很多方面都得到了广泛的应用[8]。在众多不同的深度神经网络当中，对卷积神经网络的研究是最广泛的，它在特征提取方面有很大的优势，目前在恒星光谱分类领域中应用较少。本文提出基于卷积神经网络的方法对恒星光谱相邻类的数据进行分类，并与李俊峰等使用的 BP 算法[9]进行了对比。

1 基本原理

卷积神经网络(convolutional neural network，CNN)是当前应用最广的深度学习算法，通常包括数据输入层、卷积层、下采样层、全连接层、输出层等基本层。

数据输入层主要作用是把初始数据做预处理，主要方法包括 0 均值标准化和 min-max 标准化法。本文使用 min-max 标准化对均匀采样后的光谱图做预处理，转化公式为

$$x^* = \frac{x - \min}{\max - \min} \tag{1}$$

其中，min 和 max 分别为数据的最小、最大值，x 为需要处理的数据，x^* 为标准化后的数据。

卷积层是 CNN 最核心的一个部分，主要进行特征提取，并能够有效地降低参数数目，卷积层计算公式为

$$x_j^{(l)} = \sum_{i \in N_j} a_i^{(l-1)} k_{ij}^{(l)} + b_C^{(l)}$$
$$a_i^{(l)} = f(x_j^{(l)}) \tag{2}$$

下采样层夹在连续的卷积层中间，主要作用是进行特征压缩，减小过拟合。下采样操作一般有两类，一类是最大池化，一类是平均池化。本文使用的是在实际操作中被广泛应用的最大池化法。

全连接层在卷积神经网络尾部，在整个卷积神经网络中起到“分类器”的作用。

卷积层和全连接层需要用激活函数对其进行处理。sigmoid，tanh 和 ReLU 等激活函数由于其良好的效果得到了广泛的应用。本文卷积层使用 ReLU 函数进行处理，它具有计算简单，收敛快的特点；输出层使用 sigmoid 进行激活，具有单调连续，优化稳定的特点。ReLU 激活函数的表达式为

$$f(x_j^{(l)}) = \max(0, x_j^{(l)}) \tag{3}$$

sigmoid 激活函数的表达式为

$$f(x_j^{(l)}) = \frac{e^{x_j^{(l)}}}{\sum_i e^{x_i^{(l)}}} \tag{4}$$

如图 1 所示，激励函数作用于神经网络。

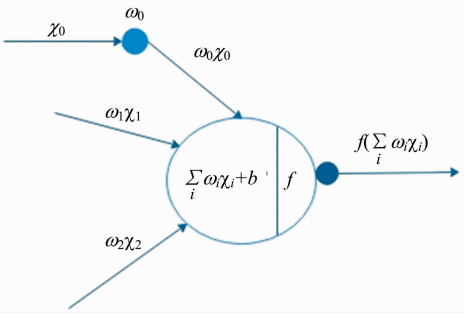

图 1 激活函数作用示意图

Fig. 1 Schematic diagram of activation function

本文使用的 CNN 结构图如图 2 所示。

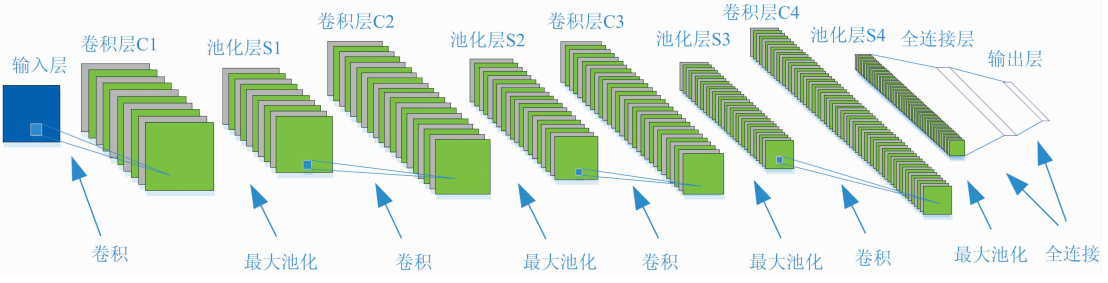

图 2 恒星光谱分类卷积神经网络结构

Fig. 2 The structure of convolutional neural network

此外为了避免过拟合并提高模型对未知数据的预测能力，在全连接层和输出层之间添加 Dropout[10-11]。它的基本思想是当网络模型进行训练时，把一些神经网络节点按一定的比例丢弃，使其暂时不发挥作用。Dropout 可以理解成是一种十分高效的神经网络模型平均方法，由于每次训练时都是随机选择不同的节点进行隐藏，使模型的鲁棒性得到了提高。

2 实验分析与讨论

实验在由 Tensorflow 搭建的框架下进行，使用的语言是 Python3．5，使用的一维光谱图是在 LAMOST 的 DR3 光谱

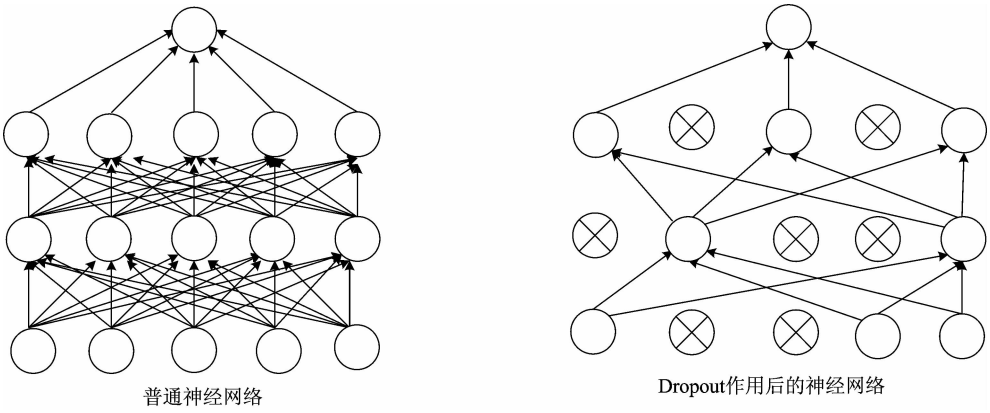

图 3　Dropout 作用前后对比图

Fig. 3　Dropout pre and post contrast diagram

库中下载。实验主要分为两部分，首先对除去 O 型星以外的六类光谱进行分类，然后对每相邻的两类光谱进行分类。预处理截取光谱 3 600～7 300 Å 的部分，然后对其均匀抽样得到所需的样本数据。对相邻类别的恒星进行分类时采用 5 折交叉验证来验证分类器的性能，即将数据集平均分为 5 份，每次实验选取一份当作测试样本，剩余部分当作训练样本，共进行 5 次实验，然后求得准确率的均值，这个方法充分利用了所有样本。实验所用各类光谱数量在表 1 和表 2 中列出。

表 1　分六类的光谱数据

Table 1　Six kinds of spectral data

光谱类型	数据集	训练集	测试集
B	1 400	1 000	400
A	1 400	1 000	400
F	1 400	1 000	400
G	1 400	1 000	400
K	1 400	1 000	400
M	1 400	1 000	400

表 2　分相邻类时的光谱数据

Table 2　Spectral data of phase dividing adjacent class

训练集/训练数据	光谱类型	数据集	训练集	测试集
OB	O	250	200	50
	B	4 000	3 200	800
BA	B	4 000	3 200	800
	A	4 000	3 200	800
AF	A	4 000	3 200	800
	F	4 000	3 200	800
FG	F	4 000	3 200	800
	G	4 000	3 200	800
GK	G	4 000	3 200	800
	K	4 000	3 200	800
KM	K	4 000	3 200	800
	M	4 000	3 200	800

首先采用 min-max 归一化方法对数据进行预处理，然后将训练集放入图 2 所示的网络进行训练和测试，迭代 3 000 次。进行迭代时各类恒星光谱分类准确率曲线如图 4 所示，可以看出随着迭代的进行，准确率趋于稳定。

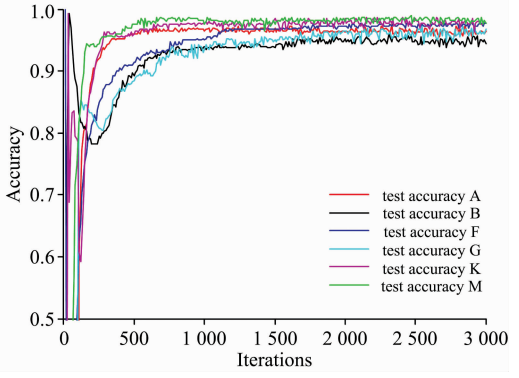

图 4　分类准确率曲线

Fig. 4　The classification accuracy curve

表 3 为对除去 O 型星以外的六类光谱进行分类的结果，可以看出实验使用的 CNN 网络对恒星光谱的分类准确率都在 95%以上，证明了 CNN 解决一维恒星光谱分类问题的可行性。

表 3　分类实验结果

Table 3　Classification experiment result

光谱类型	测试集正确分类条数	A/%
B	381	95.2
A	382	95.5
F	391	97.7
G	386	96.5
K	387	96.8
M	383	95.8

表 4 为每相邻两类光谱进行分类时的实验结果分析，其中 P 代表精确率，R 代表召回率，A 代表准确率反映了分类器对整个样本的判断能力，可以看出由于 O 型星的数据量较少，因此对 O 型星进行归类时准确率不高，其余类别光谱的

分类结果都高于 98%。本实验采用 BP 网络作为对照实验，包括 3 个隐藏层，其中各个层所包括的单元数分别是 20，40 和 20，实验结果可得 CNN 的分类准确率高于 BP 网络。

表 4 分类实验及结果分析

Table 4 Classification experiment and result analysis

光谱类型	测试集正确分类条数	P/%	R/%	F-score/%	A/%	BP 的准确率/%
O	31	93.94	62.86	75.31	—	—
B	798	97.67	99.80	98.71	—	—
O+B	829	—	—	—	81.33	78.17
B	795	98.03	99.40	98.71	—	—
A	784	99.37	98.00	98.68	—	—
B+A	1 579	—	—	—	98.70	82.20
A	795	99.80	99.40	99.60	—	—
F	798	99.40	99.80	99.60	—	—
A+F	1 593	—	—	—	99.60	84.40
F	787	99.80	98.40	99.10	—	—
G	798	98.42	99.80	99.11	—	—
F+G	1 585	—	—	—	99.10	94.20
G	790	99.75	98.80	99.27	—	—
K	798	98.76	99.80	99.28	—	—
G+K	1 588	—	—	—	99.30	91.60
K	787	98.62	98.40	98.51	—	—
M	789	98.38	98.60	98.49	—	—
K+M	1 576	—	—	—	98.50	87.31

3 结 论

使用了一种基于卷积神经网络的方法对恒星光谱进行分类。首先对除去 O 型星以外的六类光谱进行分类，每类样本的训练集分别包括 1 000 条数据，每类样本的测试集分别为 400 条光谱，在对相邻两类的恒星光谱进行分类时，将光谱数据平均分为 5 份，然后采用 5 折交叉验证法求得分类的准确率。CNN 网络的性能通过四个指标来验证：精确率 P、召回率 R、F-score、准确率 A。从实验结果可以看出 CNN 算法能够精准的对恒星光谱进行分类，在对六类恒星光谱进行分类时其准确率都在 95%以上，在对相邻类别的恒星进行分类时，由于 O 型星样本量较少，所以得到的分类结果不太理想，对其余类别的恒星分类准确率都高于 98%，通过与 BP 算法进行对比，可以看出 CNN 算法恒星光谱分类的准确率更高。

References

[1] Li Xiangru, Pan Ruyang, Duan Fuqing. Research in Astronomy and Astrophysics, 2017, 17(4): 36.

[2] Morgan W W. Chicago ILL the University of Chicago Press, 1943, 1: 3.

[3] Schierscher F, Paunzen E. Astronomische Nachrichten, 2011, 332(6): 597.

[4] Singh H P, Gulati R K, Gupta R. Monthly Notices of the Royal Astronomical Society, 1998, 295(2): 312.

[5] Daniel S F, Connolly A, Schneider J, et al. Astronomical Journal, 2011, 142(6): 203.

[6] Liu C, Cui W Y, Zhang B, et al. Research in Astronomy and Astrophysics, 2015, 15(8): 1137.

[7] XUE Jian-qiao, LI Qi-bin, ZHAO Yong-heng(薛建桥，李启斌，赵永恒). Acta Astrophysica Sinica(天体物理学报), 2000, 20(4): 437.

[8] ZHOU Jun-yu, ZHAO Yan-ming(周俊宇，赵艳明). Computer Engineering and Applications(计算机工程与应用), 2017, 53(13): 34.

[9] LI Jun-feng, WANG Yue-le, HU Sheng(李俊峰，汪月乐，胡 升). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2016, 36(10): 3261.

[10] Hinton G E, Srivastava N, Krizhevsky A, et al. arXiv preprint arXiv: 2012, 1207.0580.

[11] Hinton G E, Srivastava N, Krizhevsky A, et al. Computer Science, 2012, 3(4): 212.

Fast Classification Method of Star Spectra Data Based on Convolutional Neural Network

WANG Nan-nan[1], QIU Bo[1*], MA Jie[1*], SHI Chao-jun[1], SONG Tao[1], GUO Ping[2*]

1. School of Electronics and Information Engineering, Hebei University of Technology, Tianjin 300401, China

2. School of Systems Science, Beijing Normal University, Beijing 100875, China

Abstract Classification of stellar spectral data is one of the most basic tasks in automatic recognition of celestial spectra. The study of spectral classification can provide clues to the evolution of stars. With the development of science and technology, astronomical data are also moving towards the era of big data. The number of stars that need to be processed is increasing. How to classify them automatically and accurately has become one of the difficult problems that astronomers have to solve. At present, there are few methods to solve the problem of Star automatic classification. In this paper, a convolution neural network based method is used to classify star spectral MK system. The network is composed of data input layer, four convolution layers, four pooling layers, full connection layer and output layer. Compared with traditional network, it has the advantages of local perception and parameter sharing. In this paper, a simple and efficient convolution neural network with four convolution layers is constructed by Tensorflow in Python 3.5 environment. Dropout is applied to the full connection layer to prevent over fitting. Dropout's basic idea: When the network model is trained, some neural network nodes are discarded in a certain proportion, so that they do not play a role temporarily. Dropout can be understood as a very efficient neural network model averaging method, because it does not depend on some local features, it can make the network model more robust. The one-dimensional star spectrogram used in the experiment was downloaded from the LAMOST DR3 database. First, the spectrum was intercepted by pretreatment. After uniform sampling, it was initialized by min-max standardization method. The experiment consists of two parts. The first part classifies the spectrum according to the star spectrum MK system. Each training sample contains 1 000 spectral data and 400 spectral data. First, the CNN network is trained by training samples, and then 3 000 iterations are carried out. Then, the test samples are divided into several parts by the trained network. The second part is the classification of adjacent two types of star spectra, in which the O-type star data set sample is 250 spectra, and the rest are 4 000 spectra. The data are divided into five parts, one of which is selected as test set each time, the rest as training set, using 5 fold crossover. The accuracy of the model was calculated by the verification method, and the BP neural network was used for comparative experiments. The indicators to evaluate the network model include accuracy rate P, recall rate R, F-score and accuracy rate A. The experimental results show that the classification accuracy of the six types of stars is more than 95%. When classifying the adjacent types of stars, the classification results are not ideal because of the small sample size of O type stars. The classification accuracy of the other types of stars is higher than 98%. All the above results prove that CNN algorithm can classify the stars. The classification of stellar spectra is well solved.

Keywords Stellar spectral data; Automatic classification; CNN; 5-Cross-validation

(Received Sep. 5, 2018; accepted Jan. 19, 2019)

* Corresponding authors