基于卷积神经网络的天体光谱分类模型

【天文数据算法挑战赛 第三名CNS】

沈琢乔ouc16020021031@gmail.com

中国海洋大学，山东青岛

【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天体光谱进行分类。

【关键词】卷积神经网络 天体光谱 深度学习 LAMOST

【正文】

1. 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。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。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。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。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天文光谱数据进行分类。

1. **数据探索与方案设计**

对数据进行初步探索，绘制出图像，每条数据是2600维的频谱数据，每种类型的数据绘制的图像应当有相似的轮廓。观察数据分布发现，每条数据的振幅大小差异较大，部分数据相邻特征还有一个跃变，可以认为该数据在采集或修正过程中产生了噪声。

数据本身的振幅应该对类别的影响较小，故把数据当做2600维的频谱图，假定每条数据的总能量大小相等，应对数据进行归一化处理。



为修正数据噪声，对每条数据进行滤波处理，选用hampel滤波函数进行尝试。

既然假定每条数据是总能量相等的2600维的频谱图，则尝试选用一维卷积神经网络进行实验建模。

1. **实验与试错**

首先不对数据进行任何处理，构建一个具有多层卷积层和池化层的一维CNN的baseline，以F1-macro为mrtric，categorical\_crossentropy为损失函数，采用RAdam优化器，精度较好，显然模型匹配任务。

其次对数据进行归一化处理，模型有着不小的提升，F1-macro达到0.98。

再次尝试构建更深且更大的模型，尝试使用resnet模型并merge多个不同卷积核大小的模型。实验发现resnet模型能够构建更深的模型并达到更高的精度。训练多个不同卷积核大小的模型并merge耗时极长且没有显著提升，故不予采用。

由于观察到数据奇怪的跃变，尝试使用hampel滤波函数对数据进行修正再归一化，结果不理想，也许CNN模型自身具有较强的噪声过滤能力，滤波后损失的信息反而比过滤的噪声带来的影响更大（或许是滤波器设计不合理），故不予采用。

1. **结果与讨论**

最后采用卷积核为3的单个resnet模型，对数据进行了归一化，采用10折交叉验证，以F1-macro为mrtric，以categorical\_crossentropy为损失函数，使用RAdam优化器，10轮早停策略。在最终测试集取得F1-macro=0.98618第三名的成绩。

可以尝试添加注意力机制，调参，改进滤波器，增加数据等方法或许能更进一步。