# 2021 秋国科大数据挖掘课程大作业

## 选题：天体光谱智能识别分类


程逸飞 2021280186700451

蔡逸桐 2021180186700621

李宗儒 2021180060100402

王蓝蓝 2021280155470043

[1]中国科学院大学 网络空间安全学院

[2]中国科学院大学 资源与环境学院

[3]中国科学院大学 经济与管理学院


2021 年 11 月 21 日

# 摘要

本次作业旨在使用《数据挖掘》课上所学知识，完成针对天体光谱的智能识别分类任务。本小组通过分工合作，按照数据挖掘领域典型的操作流程，针对天体光谱数据完成了包括数据预处理、特征提取、模型构建等一系列工作。通过综合对比不同机器学习和深度学习模型的任务准确率以及时间复杂度，最终选定使用随机森林分类器构建此任务的分类模型，并在测试集上取得了 **0.9581** 的分类准确率。

# 目录

# 一． 任务背景及工作简介

天体光谱学是天文学使用的光谱学技术。研究天体的电磁辐射光谱，包括可见光，是来自恒星和其它天体的辐射。光谱学可以用来推出远距离恒星和星系的许多性质，例如它们的温度、化学组成、金属丰度，也可以从多普勒红移测量它们的运动。

天体光谱分析一般有两种：① 定性分析，用来确定天体的化学成分。首先测定谱线的波长。在拍摄天体光谱后，挡住用来拍摄天体光谱的那部分狭缝，将已知谱线波长的光源投在狭缝的其他部分上，拍摄比较光谱(常是铁弧光谱)。用仪器将天体谱线波长和地球上已知元素的谱线波长作比较，或者应用按原子结构和光谱理论计算的谱线表，证认出产生天体谱线的元素。② 定量分析，每种元素的谱线强度,与它们在物质中的含量有关，所以通过对谱线强度的比较，可以确定物质中各元素的含量。对于天体，目前只能取到月球上的物质样品，在实验室中进行定量分析。至于恒星（包括太阳）光谱的定量分析，有两种方法：一是测定一些谱线的等值宽度，作出观测的生长曲线，与理论计算比较；二是根据某种谱线形成的机理，假设一些物理参数，计算出理论轮廓，再同观测轮廓比较。这两种方法不仅能得到形成该谱线元素的原子数，而且能得到恒星大气中的温度、湍流速度和压力等参数。

本项任务中要求对恒星、星云和类星体光谱进行分类识别，这三类光谱在天文物理学的研究上是有一定区分特征的。恒星光谱中的吸收线可以用于确定恒星的化学成分；星云中原子的行为和被压抑的谱线与在正常密度下非常的不同。这些谱线是所谓的禁线，并且是星云光谱中最强的谱线。类星体的光谱类似于普通的星系光谱，但被认为有着高度红移。

本项团队工作旨在通过数据挖掘的方法对三种天体进行光谱层面的智能识别分类。工作主要分为三大部分：数据预处理、使用机器学习模型分类和使用深度学习模型分类。具体而言，在数据预处理阶段我们首先对数据样本进行数据清洗，包括去除异常点、标准化和归一化等，同时面向数据不平衡问题提出了两种解决方案即过采样配平法和增加类别平衡权重法。接下来对于机器学习模型我们在预处理阶段还增加了特征提取的工作以便于机器学习算法能够更好更快的学习样本特征。数据预处理阶段结束后，分别训练机器学习模型和深度学习模型完成最终的分类任务。其中机器学习算法采用了随机森林和支持向量机进行对照试验并结合特征提取和类别平衡权重法来增强模型的训练效果；深度学习部分采用卷积神经网络和 50 层 ResNet[1]网络进行对照试验，同时使用过采样配平法缓解数据类别不平衡问题。本工作的整体处理流程如图 1 所示。

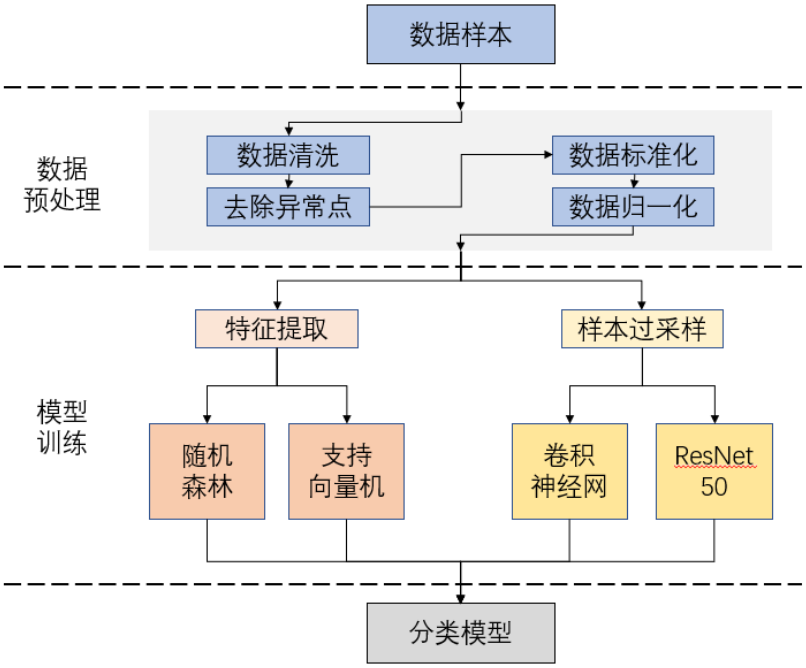


图 1 任务流程示意图

# 二． 数据预处理

## 2.1 数据预览

本次大作业是一个天体光谱智能识别分类的任务，数据集中共 18 万条样本数据，每个样本有 2600 维。整个样本空间一共包含三类数据分别是行星（star），星系（galaxy）和类星体（qso）。要做的是根据样本数据进行多分类，评价指标是分类准确率。

首先从考察任务整体难易程度出发，直接观察多分类任务中各类别之间的区分程度，这里为方便展示，设计两个实验使用余弦相似度观察类别之间的可区分程度并使用热力图进行可视化展示，实验一随机抽取三个类别中一个样本计算余弦相似度，实验二为进一步消除偶然误差影响随机抽取每个类别中的 10 个样本求平均值后计算余弦相似度。实验一二的可视化结果如图 2 所示。

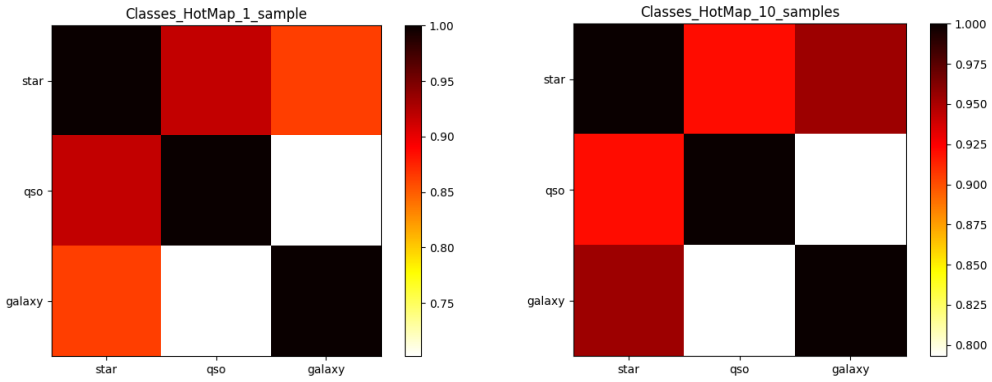


图 2 样本相似度热力图

由上图可以看出，对角线代表的是自相关系数，其相似性始终为 1 且颜色最深，另外值得注意的是，star 类对于其他两类的相似性都比较高，这表明样本数据不是非常好区分；qso 类和 galaxy 类的相似性矩阵表明这两类的区分程度很强，对模型来说比较友好。

## 2.2 缺失值处理

我们检测数据中是否有空值等影响分类的数值，结果显示数据集质量较高不存在上述问题。

## 2.3 异常点处理

通过直接对某个维度的数据分布情况进行考察，发现在异常点的影响下导致数据分布跨度很大，这里以维度一为例，对该维度下的取值数据进行排序后绘制散点图观察如图 3，可以看到大部分数据都集中分布在正负 1e6 附近，但有异常点出现在了负 1e7 尺度内，导致这是直接做数据归一化的话，会将绝大部分数据归一化到 0 附近，这是相当没有意义的，因此十分有必要去除异常点之后再进行归一化。

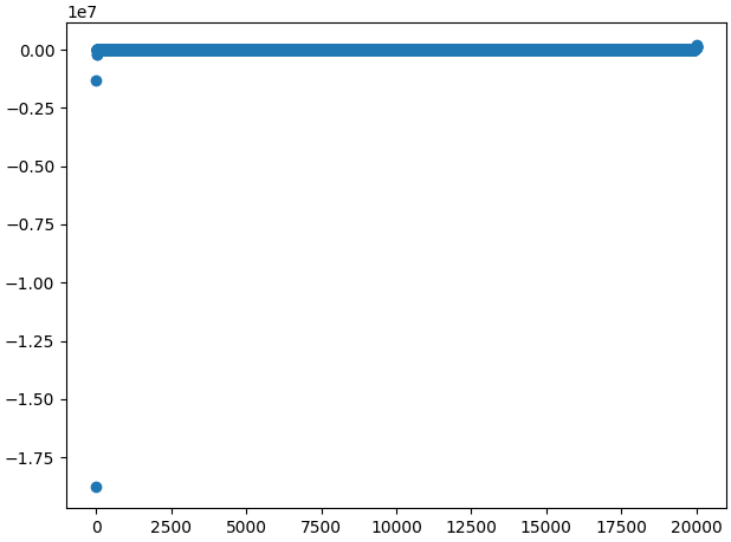


图 3 数据分布散点图

使用均值加减三倍标准差检测异常点，以文件一为例，对样本的的每一维特征单独提取并计算每一个样本数据在该特征下是否超过均值加减三倍标准差即是否为异常点，若为异常点则在该维度下对该样本进行标记，最后将共 2600 维分别的异常点标记进行取集合运算就能得到要提出的异常点了。如上所述操作后，共 20000 条样本，对 2600 维分别检测异常点后共得到了 6796 个异常样本（有重复的），对这些样本取集合运算得到共 139 条样本，考虑到本任务样本充足，

故在对类别平衡影响不大（异常样本全是 star 类，属于多数类）的情况下直接将异常样本进行删除处理。

最终对于共 9 个文件的剔除的异常点分布情况如下表 1。可以看出剔除的异常点绝大部分都是 star 类别的数据，而整个数据集恰好存在 star 类别数量过多的数据不平衡的问题，因此直接剔除检测出的这些异常点是无可厚非的。

| | star 数量 | qso 数量 | galaxy 数量 |
|---|---|---|---|
| 文件一（训练集） | 139 | 0 | 0 |
| 文件二（训练集） | 532 | 0 | 1 |
| 文件三（训练集） | 403 | 1 | 1 |
| 文件四（训练集） | 128 | 0 | 1 |
| 文件五（训练集） | 399 | 0 | 1 |
| 文件六（训练集） | 66 | 0 | 1 |
| 文件七（训练集） | 427 | 0 | 1 |
| 文件八（验证集） | 21 | 0 | 0 |
| 文件九（验证集） | 213 | 0 | 2 |

表 1 异常点去除情况统计

这里以数据样本的第一个特征进行展示，观察数据分布情况。图 4 为按照该特征提取样本之后，进行数据排序之后绘制散点图。观察发现大部分数据呈类似反函数形式分布在空间内，也有一部分点出现在负空间内，但总体样本分布较为均匀不再出现异常点。

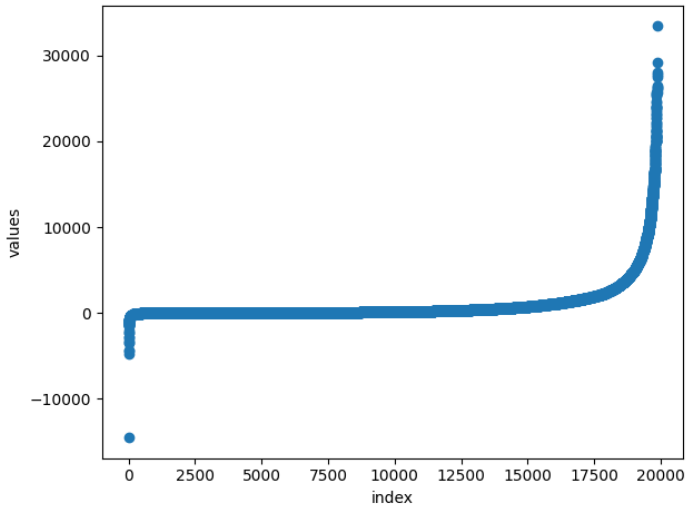


图 4 数据分布散点图

其次依然观察该数据的概率密度分布图（图 5），发现数据呈现尖峰重尾特性，具有长尾分布的特征，符合常理。

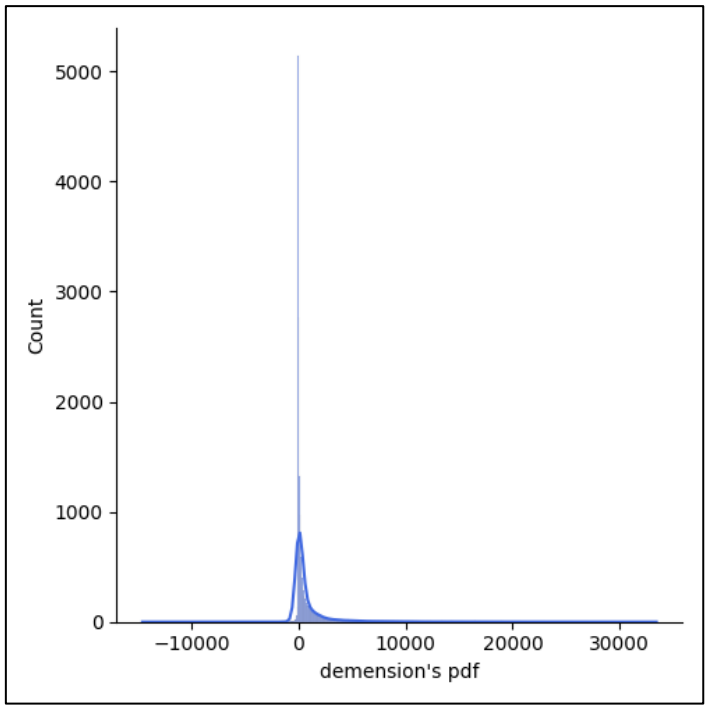


图 5 数据分布概率密度图

## 2.4 标准化与归一化

从上一步结果得到的数据分布我们可以看出，特征的数据分布仍然是存在度量偏差的。使用 z-score 标准化[3]（公式 2-1）方法能将样本数据的每一维特征的转化为标准正态分布。另外为了进一步消除不同特征之间的量纲影响，再使用 minmax 归一化（公式 2-2）将数据约束在[0,1]之间。

$$\gamma' = \frac{\gamma - \mu_A}{\sigma_A} \qquad \text{公式 2-1}$$

$$\gamma' = \frac{\gamma - min_A}{max_A - min_A} \qquad \text{公式 2-2}$$

## 2.5 特征提取

原始数据样本有 2600 个特征维度，这些特征一方面可能存在无用甚至干扰分类的特征，另一方面对于机器学习模型的训练也将带来很大的开销。因此，在进行机器学习模型进行分类任务之前，我们准备对数据进行降维处理。我们考虑

了 PCA 和 tSNE 两种特征提取（降维）的方法。

为了比较两种方法特征提取的好坏，我们直接使用原始分类任务进行特征提取的评价，这里为了节约运算资源，只对前两万个数据样本进行实验，用两种算法特征提取结束之后加入随机森林分类器，因为特征提取的好坏是要以最终任务目标为导向的，所以我们以分类任务的准确性判断特征提取的好坏。使用 PCA 将数据特征维度降低到 200 维。由于 tSNE 相对于 PCA 计算的时间成本高很多，且常见的 tSNE 多用在数据可视化场景下，因此我们对于 tSNE 直接将数据降到 3 维，且为加快计算我们采用 PCA 的 200 维结果进行 tSNE 进一步降维。特征提取工作的测试结果见表 2。

| | 最终维度 | 得分（准确率） | 耗时 |
|---|---|---|---|
| PCA | 200 | 0.9504 | 1（设为单位时间） |
| PCA+tSNE | 3 | 0.9514 | 45 |
| PCA+tSNE | 2 | 0.9498 | 30 |

**表 2 降维结果分析表**

我们知道 PCA 侧重于将数据样本重构表示，而 tSNE 侧重于将数据样本按类别尽可能区分表示，因此 PCA+tSNE 的处理方式得到的结果虽然最后只有 3 维却也能很好地描述整个数据集的数据分布情况。下图 6 是使用 PCA+tSNE 降维后的可视化图像，可以看到经过这种降维后数据分布情况较为清晰，有利于分类模型完成分类任务。

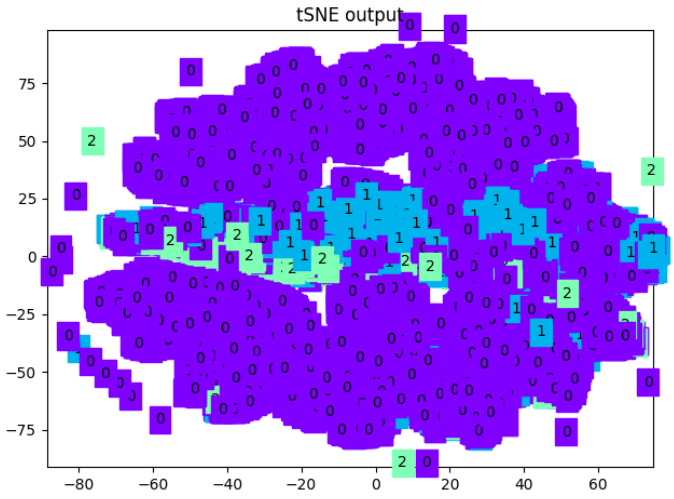


图 6 tSNE 降维结果可视化图

## 2.6 数据样本过采样

由于数据样本类别数量极不平衡，star、galaxy 和 qso 三类样本所占数据集比例为 83.93：12.27:3.80，因此由数据不平衡问题将引起比较严重的训练误差，导致对于较多数类训练次数更多同时降低了对少数类分类正确性的要求。但根据我们对多分类任务的普遍认识，如果没有特殊要求，各类分类准确性应该是同等重要的，因此我们需要解决数据不平衡问题。

处理数据不平衡问题的方法主要有两类，其一是将数据样本少数类进行过采样、多数类进行欠采样；其二是训练过程中对于评价指标增加类别平衡权重，简单理解就是更加看重少数类的分类准确性。我们本次实验同时实施这两种方法，对于使用机器学习算法的部分采用增加类别权重的方法，对于使用深度学习的部分采用数据过采样配平技术。这里的过采样我们采用 SMOTE 算法即平衡少数类算法，其相对于随机过采样产生的数据更加贴近真实。平衡后训练集三类样本数量均为 110891。

# 三． 模型方法论

本项工作中主要使用两项机器学习算法（随机森林和支持向量机）和两项深度学习算法（卷积神经网络和残差神经网络 ResNet）进行对照试验，下面将依次对算法实现进行描述。

## 3.1 随机森林分类模型

随机森林算法[2]是一种集成学习算法，集成学习的思想是为了解决单一模型（弱学习器）进行训练测试时固有的缺陷，整合多个模型进行取长补短以避免局限性。随机森林就是基于这种思想，通过整合多个 CART 决策树来组成森林构成更加强大的分类器。接下来将首先介绍其中的弱学习器----CART 决策树。

需要注意的是朴素观点的决策树一般有三种，它们分别是使用信息增益选择特征的 ID3 决策树、使用信息增益比选择特征的 C4.5 决策树和使用基尼指数选择特征的 CART 决策树。对于 CART 决策树来说，使用的基尼指数代表了模型的不纯度，基尼指数越小则不纯度越低，其对应的特征越好，这和信息增益以及信息增益比恰好相反。设基尼指数使用$Gini()$表示，有 K 个类别，取第 k 个类别的概率为$p_k$，则概率分布的基尼指数表达式如公式 3-1 所示。

$$Gini(p) = \sum_{k=1}^{K} p_k\,(1 - p_k) = 1 - \sum_{k=1}^{K} {p_k}^2 \qquad \text{公式 3-2}$$

进一步地，对于样本$D$,数量为$|D|$，假设有 K 个类别，第 k 个类别的数量为$|k|$，则有样本 D 的基尼指数为公式 3-2 所示的表达式。

$$Gini(D) = 1 - \sum_{k=1}^{K} (\frac{|k|}{|D|})^2 \qquad \text{公式 3-2}$$

将特征考虑进去，依然对于数量为$|D|$的样本$D$，根据特征$A$的某个值$a$可以将样本$D$分成$D_1$和$D_2$，则在此特征$A$的条件下，样本$D$的基尼指数表达式为公式 3-3 所示。

$$Gini(D,A)=\frac{|D_1|}{|D|}Gini(D_1)+\frac{|D_2|}{|D|}Gini(D_2) \qquad \text{公式 3-3}$$

有了如上基尼指数的公式，就可以根据下面的伪代码流程生成 CART 决策树:

**算法 2** CART 决策树算法

**输入：**训练集$D$，基尼系数阈值$gini$，样本个数阈值$c$

**输出：**CART 决策树 Tree

1: 算法从树的根节点开始，通过训练集递归生成 Tree

2: 对于当前节点的节点数据集$D$，如果样本数量小于阈值$c$或者没有可用特征，则返回决策子树并且停止当前节点递归

3: 对于当前节点计算样本集$D$的基尼指数，如果小于$gini$则返回决策子树并且停止当前节点递归

4: 计算当前节点现有特征的每个特征值对数据集$D$的基尼指数

5: 在步骤 4 计算出的结果中选择基尼指数最小的特征$A$和对应的特征值$a$，使用该特征值将数据集划分成$D_1$和$D_2$，同时建立当前节点的左右节点并使用数据$D_1$和$D_2$作为孩子节点的数据集

6: 对左右的孩子节点递归调用步骤 2-5，生成决策树 Tree

在介绍完随机森林算法使用的弱学习器之后，随机森林算法即使用自助抽样集成方法（Bagging）进行集成。自助抽样集成首先将训练集随机划分成 m 个新的数据集，而后在每一个数据集上构建一个弱学习器模型----此处是一个 CART 决策树，它们是相互独立地进行训练，最后预测的时候对这 m 个模型进行整合获得最终的结果。此处随机森林算法解决的是分类问题，最终结果使用弱学习器投票法，取投票的多数类作为模型分类的结果。

## 3.2 支持向量机分类模型

对于分类问题,支持向量机算法根据区域中的样本计算该区域的决策曲面，由该曲面确定该区域中样本的类别。样本 x 为 k 维向量，在某区域内的 l 个样本所属类别为 (x1 , y1 ),...,(xl , yl )∈ $R_k \times \{\pm 1\}$ 。若超平面:

$$\omega \cdot x + b = 0 \qquad \text{公式 3-4}$$

能将样本分为两类，其中×表示向量的点积。最佳的超平面应使两类样本到超平面的距离为最大。显然，公式 3-4 中的 w 和 b 乘以系数以后仍能满足方程。不失一般性，对于所有的样本 xi ，式 w× xi + b 的最小值为 1。则样本与此最佳超平面的最小距离为 |(w× xi + b)|/||w|| =1/||w|| 。最佳超平面应满足约束:

$$y_i[(\omega \cdot x) + b] > 1, i = 1, \dots, l \qquad \text{公式 3-5}$$

w 和 b 的优化条件应该是使两类样本到超平面最小距离之和 2/||w||最大。另外,考虑到可能存在一些样本不能被超平面正确分类,因此引入松弛变量:

$$\vartheta_i \geq 0, i = 1, \dots, l \qquad \text{公式 3-6}$$

问题变成在公式 3-5 和公式 3-6 的条件下最小化

$$\frac{1}{2}||\omega||^2 + C \cdot \sum_{i=1}^{l} \vartheta_i \qquad \text{公式 3-7}$$

其中 C 为一个正常数。上式的第 1 项使样本到超平面的距离尽量大，从而提高泛化能力；第 2 项则使误差尽量小。利用 lagrange 乘子法，可以把公式 3-7 变成其对偶形式，从而有

$$\begin{aligned} & maxW(a) = \sum_{i=1}^{l} a - \frac{1}{2}\sum a_i y_i a_j y_j (x_i \cdot x_j), \\ & s.t. \qquad \sum_{i=1}^{l} a_i y_i = 0, \\ & a_i \in [0, C], i = 1, \dots, l \end{aligned} \qquad \text{公式 3-8}$$

以及

$$w = \sum_{i=1}^{l} a_i y_i x_i \qquad \text{公式 3-9}$$

这是一个典型的二次优化问题，已由高效的算法求解。可以证明，在此优化问题的解中有一部分 ai 不为 0，它们所对应的训练样本完全确定了这个超平面，因此称其为支持向量.按照优化理论的 Kuhn-Tucker 定理，在鞍点，对偶变量与约束的乘积为 0，从而求得超平面的另一个参数 b 满足：

$$y_i(w \cdot x_i + b) = 1 \qquad \text{公式 3-10}$$

对于未知属类的向量 x，可以采用线性判决函数

$$f(x) = sgn(w \cdot x + b) \qquad \text{公式 3-11}$$

来判定其所属类别。综合公式 3-9，得到：

$$f(x) = sgn(\sum_{i=1}^{l} a_i y_i (x_i \cdot x) + b) \qquad \text{公式 3-12}$$

## 3.3 卷积神经网络

卷积神经网络是一种深度前馈神经网络，它一般包括有卷积操作、池化操作和全连接操作。这种网络模型具有表征学习的能力并且可以根据输入数据的层次结构对其进行准确地分类。本文使用的孪生网络模型中的基础组成部分就是两个共享神经元和权重的卷积神经网络，另外本文的模型比较阶段也使用该网络进行异常行为检测任务比较。随着计算机硬件性能的不断迭代升级，计算能力得到了很大的提升，这使得卷积神经网络能够完成更快速的训练和测试。另外，卷积神经网络的隐含层能够进行卷积核参数的共享，加之层间连接的稀疏性，使得减少计算量的同时也能有效提取有价值的信息成为可能。

卷积神经网络的核心是卷积层和池化层，它们对应的操作如公式 3-13 和公式 3-14 所示。使用输出的具有$l$层的特征映射（feature map）$Z^l$和第$l+1$层的卷积核$w^{l+1}$进行自相关系数的计算。当上述卷积层的输出形成了之后，紧接着进行的是池化层操作（如果存在的话），通过池化层处理后就得到了整个第$l+1$层的

输出。在卷积层和池化层的公式中，使用$K$表示特征映射的通道数量，$f$表示相映卷积核和池化核的尺寸，$s_0$表示步长，$p$表示填充的大小。对于结构化的输入信息，每一个卷积核在输入的特征上有规律地横移以扫过每个特征，每一步都进行矩阵乘法并叠加特征中的偏差值。使用多种卷积核能够从不同的角度提取到高阶的全局特征信息。池化操作能够完成特征选择和信息过滤的工作，对卷积核的输出进行欠采样选择。根据预先设置的池化操作，池化层能够将某一个点特征转化到区域特征，以此来进一步聚合特征特性并能够降低模型的过拟合程度。

$$Z^{l+1}(i,j) = [Z^l \otimes w^{l+1}](i,j) + b$$
$$= \sum_{k=1}^{K_l} \sum_{x=1,y=1}^{f} [Z_k^l(s_0 i + x, s_0 j + y) w_k^{l+1}(x,y)] + b \quad \text{公式 3-13}$$

$$Z_k^l(i,j) = [\sum_{x=1,y=1}^{f} Z_k^l(s_0 i + x, s_0 j + y)^p]^{\frac{1}{p}} \quad \text{公式 3-14}$$

本实验中使用的卷积神经网络结构如图 7 所示。

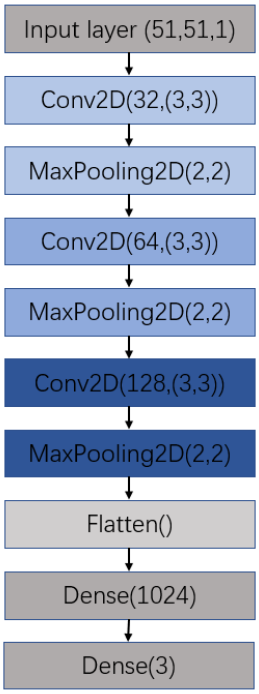


图 7 CNN 结构图

## 3.4 残差神经网络 ResNet

ResNet 的主要思想是在网络中增加了直连通道，即 Highway Network 的思想。此前的网络结构是性能输入做一个非线性变换，而 Highway Network 则允许保留之前网络层的一定比例的输出。ResNet 的思想和 Highway Network 的思想也非常类似，允许原始输入信息直接传到后面的层中，如下图 8 所示。

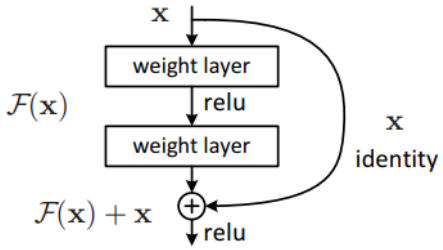


图 8 残差学习模块

提出残差学习的思想。传统的卷积网络或者全连接网络在信息传递的时候或多或少会存在信息丢失，损耗等问题，同时还有导致梯度消失或者梯度爆炸，导致很深的网络无法训练。ResNet 在一定程度上解决了这个问题，通过直接将输入信息绕道传到输出，保护信息的完整性，整个网络只需要学习输入、输出差别的那一部分，简化学习目标和难度。本实验采用的 ResNet50 结构如图 9 所示。

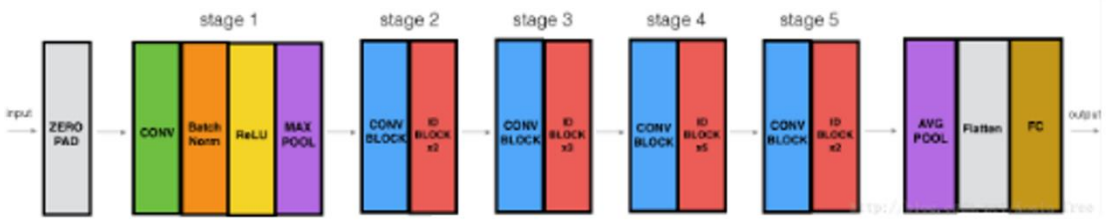


图 9 ResNet50 结构图

# 四． 实验结果

## 4.1 评价度量指标

本文将使用准确率（Accuracy)、精确率（Precision）和召回率（Recall）作为评价模型检测能力的主要指标，它们的计算方式分别参考公式 4-1 和公式 4-2。$F_\beta$是另一个非常重要的评价指标（公式 4-3），它通过结合精确率和召回率这两个子指标来生成新的指标，以此表征更加综合的模型检测能力。在$F_\beta$中，参数$\beta$是用来非平衡考虑精准率和召回率的权重值，$\beta$越大（大于 1）表示评价过程中召回率更重要；$\beta$越小（小于 1）表示评价过程中精确率更重要。在本文中，精确率和召回率对于模型检测评价同样重要，故设置$\beta$为 1。

$$Precision(P) = \frac{TP}{TP+FP} \quad \text{公式 4-1}$$

$$Recall(R) = \frac{TP}{TP+FN} \quad \text{公式 4-2}$$

$$F_\beta = \frac{(1+\beta^2)\times P\times R}{(\beta^2\times P)+R} \quad \text{公式 4-3}$$

## 4.2 模型训练及结果分析

首先对于两个机器学习算法—随机森林和支持向量机进行分类任务的构建，这里设置类别平衡权重来缓解数据不平衡的问题，同时在训练集上进行五折交叉验证，训练过程中的指标评价见表 3。

| | Precision_macro | Recall_macro | F1_macro | Time_consuming |
|---|---|---|---|---|
| RF | **0.9275** | **0.8066** | **0.8585** | **1**(controlled) |
| SVM | 0.4860 | 0.7082 | 0.4991 | 80.32 |

**表** 3 RF VS SVM

从上表可以看出，对于该项分类任务的完成上，随机森林分类器无论是准确率还是召回率都明显优于支持向量机分类器，这可能是由于随机森林集成了很多 CART 决策树分类器并进行投票带来的优势，而支持向量机只能训练出一些超平

面进行分类，其模型拟合能力相对较差。另外随机森林分类器在时间消耗上也明显优于支持向量机分类器。

将训练好的模型应用到测试集上，结果也不出我们所料，随机森林分类器获得了 **0.9581** 的准确率，而支持向量机分类器只获得了 **0.7064** 的分类准确率。两个模型在测试集上分类结果的混淆矩阵参见表 4 和表 5。

| RF Confusion_matrix | | Prediction | | |
|---|---|---|---|---|
| | | star | galaxy | qso |
| Reference | star | **33245** | 57 | 39 |
| | galaxy | 960 | **3850** | 91 |
| | qso | 434 | 103 | **985** |

**表 4 随机森林测试集结果混淆矩阵**

| SVM Confusion_matrix | | Prediction | | |
|---|---|---|---|---|
| | | star | galaxy | qso |
| Reference | star | **24303** | 5801 | 3237 |
| | galaxy | 155 | 2311 | **2435** |
| | qso | 16 | 27 | **1479** |

**表 5 支持向量机测试集结果混淆矩阵**

接下来使用深度学习的技术完成天体光谱分类的任务。卷积神经网络最典型的用法使用来处理图片，因为它需要用卷积核来提取特征，我们这里的数据样本为 2600 维，不妨再 padding 一列使得每条样本可以转换成一个 51*51 的矩阵，于是我们对于每条样本都转换构造一个类似于灰度图的样本这样就可以直接输入进卷积神经网络了。对于 ResNet 我们同样这么做。对于卷积神经网络的训练过程如图 10 所示。对于 ResNet 的训练过程如图 11 所示。

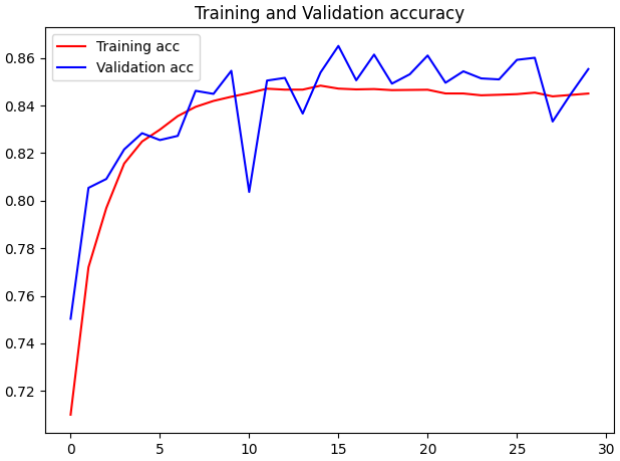


图 10 卷积神经网络训练过程

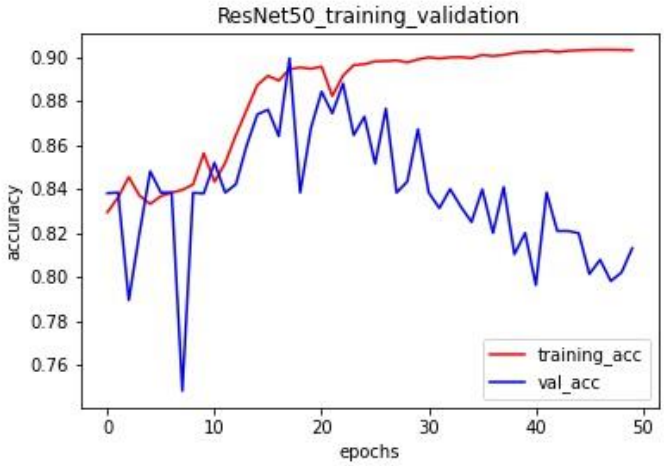


图 11 ResNet 训练过程

从训练过程可以看出卷积神经网络模型准确率最终稳定在 0.84 附近，而层数更深的 ResNet 在训练集上的准确率能够持续上升，但是验证集上准确率下降说明了模型已经过拟合了，因此在训练十五轮左右时进行了模型的保存。最终我们用训练得到的模型在测试集上测试，得到的准确率分别为 **0.8483** 和 **0.8803**。另外值得注意的是，由于 ResNet 层数很深所以相对于普通的卷积神经网络其每个 epochs 时间消耗会长很多，在 NVIDIA TITAN XP 显卡上运行本文使用的卷积神经网络训练一个 epochs 大约需要 20 秒，而 50 层的 ResNet 网络一个 epochs 大约需要 230 秒。

# 五． 总结展望

本次大作业旨在对天体光谱进行智能识别分类，通过《数据挖掘》课上所学的知识，我们进行了团队分工合作（具体任务分工见表 6），采用了一系列全流程的数据挖掘方法来完成分类工作，使用的数据挖掘技术包括数据预处理（数据清洗、异常点检测、标准化与归一化）、特征提取、样本过采样、机器学习模型（随机森林、支持向量机）和深度学习模型（卷积神经网络、残差神经网络）等。最终综合考虑模型分类能力和时间复杂度，我们训练出了任务最优的随机森林分类器，在测试集上准确率达到了 **0.9581**。

| 姓名 | 完成工作 |
| --- | --- |
| 程逸飞 | 机器学习模型及报告撰写 |
| 蔡逸桐 | 深度学习模型 |
| 李宗儒 | 特征提取与数据过采样 |
| 王蓝蓝 | 背景知识整理及数据预处理部分 |

**表 6 任务分工表**

理论上来说，深度学习模型相对于传统的机器学习应该具有更强的学习能力，因此在同等情况下应该能获得更好的分类效果，但是我们通过实验得到的结果却没有印证这一点。我们推测可能是由于过采样阶段处理有些问题或者深度学习模型参数设置不当导致了其分类性能并不理想。未来研究方向可以集中在研究形成更有利于深度学习模型训练的数据表达形式，并通过调节参数来使深度学习模型获得更好的分类效果。

# 参考文献

[1] He K， Zhang X， Ren S , et al. Deep Residual Learning for Image Recognition[J]. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[2] 周志华. 机器学习[M]. 北京：清华大学出版社, 2016

[3] Jiawei Han, Micheline Kamber, Jian Pei. Data Mining Concepts and Techniques(Third Edition)[M]. Morgan Kaufmann, 2012.