



中国科学院大学

University of Chinese Academy of Sciences

硕士学位论文

基于 Galaxy Zoo DECaLS 的棒旋星系物理特征的统计研究

作者姓名: 穆子豪

指导教师: 沈世银 研究员 中国科学院上海天文台

学位类别: 理学硕士

学科专业: 天体物理

培养单位: 中国科学院上海天文台

2023 年 6 月

**Statistical study of the physical parameters of barred spiral
galaxies based on Galaxy Zoo DECaLS**

**A thesis submitted to
University of Chinese Academy of Sciences
in partial fulfillment of the requirement
for the degree of
Master of Natural Science
in Astrophysics**

By

Zihao Mu

Supervisor: Professor Shiyin Shen

Shanghai Astronomical Observatory, Chinese Academy of Sciences

June, 2023

中国科学院大学 学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的指导下独立进行研究工作所取得的成果。承诺除文中已经注明引用的内容外，本论文不包含任何其他个人或集体享有著作权的研究成果，未在以往任何学位申请中全部或部分提交。对本论文所涉及的研究工作做出贡献的其他个人或集体，均已在文中以明确方式标明或致谢。本人完全意识到本声明的法律结果由本人承担。

作者签名：

日 期：

中国科学院大学 学位论文授权使用声明

本人完全了解并同意遵守中国科学院大学有关收集、保存和使用学位论文的规定，即中国科学院大学有权按照学术研究公开原则和保护知识产权的原则，保留并向国家指定或中国科学院指定机构送交学位论文的电子版和印刷版文件，且电子版与印刷版内容应完全相同，允许该论文被检索、查阅和借阅，公布本学位论文的全部或部分的内容，可以采用扫描、影印、缩印等复制手段以及其他法律许可的方式保存、汇编本学位论文。

涉密及延迟公开的学位论文在解密或延迟期后适用本声明。

作者签名：

日 期：

导师签名：

日 期：

摘要

盘状星系中的棒结构特征与星系自身所处环境及整体恒星形成特征之间存在复杂的相互关系。我们通过细致的星系形态特征测量计划 Galaxy Zoo DECaLS 中的数据构建了一组目前最大的具有同样恒星质量分布的强棒、弱棒和无棒星系样本。通过引入星系所处暗晕质量及相空间特征来控制星系所处的环境效应之后，进一步对比研究强棒、弱棒和无棒星系的恒星形成特征。研究表明，有棒星系中的恒星形成熄灭的比例系统的高于无棒星系，且强棒的效应更为明显。但是，对于仍处于恒星形成主序上的星系来说，强棒、弱棒和无棒星系的恒星形成特征并没有显著差别。该结果表明，星系中的棒特征在其存在的时标中可能使部分星系迅速的从恒星形成转入熄灭状态。此外，我们利用随机森林算法对星系的棒特征和其余 12 个物理参数之间存在的相关性进行回归分析，并设计了弱棒星系在物理特征上更偏向于无棒星系还是强棒星系的统计检验。基于随机森林算法给出的重要性排序，我们发现星系的轴比和 Sersic 指数与星系的强棒特征具有最强的相关性，而星系在暗晕中的投影距离特征则和弱棒特征存在显著相关性。

关键词：棒旋星系，恒星形成，环境，随机森林

Abstract

The bar structure of disk galaxies, their environment, and overall star formation properties are known to be intricately linked. By utilizing data from the Galaxy Zoo DECaLS program, we have assembled the largest sample of strong bar, weak bar, and no bar galaxies with matching stellar mass distributions. We analyzed the specific star formation rates (sSFR) of these galaxies after controlling for environmental effects by matching the halo mass and phase space diagram of the dark matter halo in which the galaxy resides. Our findings suggest that the percentage of quenched galaxies in the barred galaxy sample is consistently higher than that in no bar galaxies, with a more pronounced effect for strong bar galaxies. However, for galaxies that are still on the star formation main sequence, there is no significant difference in their sSFR distribution. These results suggest that the presence of a bar in a galaxy may rapidly transition some galaxies from being star-forming to quenched in a short period of time. In addition, we used the random forest (RF) algorithm to conduct regression analysis of the correlations between the bar features and other physical parameters of galaxies. Based on RF regression, we also designed a statistical test of whether the weak bar galaxies are more favorable to the no bar galaxies or the strong bar galaxies in terms of their physical parameters. With the importance ranking given by the RF regression, we find that the axial ratio and Sersic index have the strongest correlation with the strong bar feature of the galaxies, while the projected distance of the galaxies in their host dark halo has a significant correlation with the weak bar feature.

Key Words: Barred spiral galaxy, SFR, Environment, Random forest

目 录

第 1 章 引言	1
1.1 星系的形态及基础物理参数	1
1.2 棒旋星系	4
1.3 本文的研究目的	6
第 2 章 本文使用的数据集及样本	7
2.1 Galaxy Zoo DECaLS	7
2.1.1 Galaxy Zoo	7
2.1.2 DESI legacy survey	8
2.1.3 Galaxy Zoo DECaLS	10
2.2 NSA 数据集	14
2.3 GSWLC-X2	14
2.4 星系群数据集	15
2.5 所用样本	16
第 3 章 棒星系与恒星形成特征的统计分析	21
3.1 统计中检验样本相关性算法	21
3.1.1 K-S 检验	21
3.1.2 随机森林算法	22
3.2 棒星系分布比例	27
3.3 棒星系与恒星形成	27
3.4 棒星系与环境	30
3.5 控制环境因素后棒星系的恒星形成	32
第 4 章 棒特征与其它物理参数的关系	35
4.1 模型所用样本及物理参数	35
4.2 随机森林模型	36
4.2.1 “强棒星系”和“非强棒星系”	36
4.2.2 “有棒星系”和“无棒星系”	39
第 5 章 总结与展望	45
参考文献	47
致谢	51
作者简历及攻读学位期间发表的学术论文与其他相关学术成果	53

图目录

图 1-1 左图为 NGC 1300, 右图为 NGC 1232	2
图 1-2 哈勃序列, 也被称作哈勃音叉图 (来自 COSMOS)	2
图 1-3 三种不同的“熄灭”机制, 图源 Peng et al. (2010)	4
图 2-1 Galaxy Zoo 网站中的分类投票页面	8
图 2-2 BASS、DECaLS 和 MzLS 三个测光巡天项目天区覆盖图	9
图 2-3 GZD-1 和 GZD-2 所用的决策树	12
图 2-4 GZD-5 所用的决策树	13
图 2-5 各数据集的 AITOFF 投影覆盖范围	16
图 2-6 获取“Morph”样本所用流程图。其中判定为强棒星系需遵循“无棒投票比例 < 50% 且强棒投票比例 > 弱棒投票比例”, 判定为弱棒星系需遵循“无棒投票比例 < 50% 且强棒投票比例 < 弱棒投票比例”, 判定为无棒星系需遵循“无棒投票比例 > 50%”	17
图 3-1 决策树流程图	24
图 3-2 随机森林结构示意图	26
图 3-3 三种棒星系在不同恒星质量分布上的比例	27
图 3-4 样本“Morph-SFR”中三种棒的恒星质量和比恒星形成率 (sSFR) 的散点图和轮廓图, 其横坐标为恒星质量 ($\log(M_*/M_\odot)$), 纵坐标为比恒星形成率 (sSFR)。上边际图为恒星质量分布直方图, 右边际图为 sSFR 分布直方图。红色图例为强棒样本 (SB:5151 个), 绿色为弱棒样本 (WB:12833 个), 蓝色为无棒样本 (NB:25237 个)	28
图 3-5 图 (a) 为控制了恒星质量 (M_*) 的三种棒的比恒星形成率 (sSFR) 的直方图。图 (b) 为控制了恒星质量 (M_*) 的三种棒样本中其“Quench + GV”星系占其样本的比例的点线图。图 (c) 为控制了恒星质量 (M_*) 的三种棒中恒星形成星系部分的比恒星形成率 (sSFR) 的直方图。红色图例代表强棒, 绿色代表弱棒, 蓝色代表无棒	29
图 3-6 图 (a) 为“Morph- M_h -SFR”样本中, 控制了 M_* 的三种棒的 M_h/M_\odot 直方图。三种棒的样本个数各为 4273 个。图 (b) 为“Morph- M_h/σ -SFR”样本中, 控制了 M_* 和 M_h/M_\odot 的三种棒星系的相空间图。图中黑线为 Jaffé et al. (2015) 的方法。三种棒的样本个数各为 1031 个。红色图例代表强棒, 绿色代表弱棒, 蓝色代表无棒	31
图 3-7 左列为位力化星系样本, 右图为内落星系样本。(a)(b) 为 sSFR 的分布直方图 ((a): 各 786 个星系, (b): 各 138 个星系), (c)(d) 为不同棒样本中, 其“Quench”+ “Green Valley”星系占其样本的比例的点线图 ((c): 各 786 个星系, (d): 各 138 个星系)。(e)(f) 为三种棒星系样本中 SF 部分 sSFR 的分布直方图 ((e): 共 1503 个, (f): 共 300 个)。图中控制了恒星质量和暗晕质量	33

图 4-1 “强棒星系”和“非强棒星系”，下采样方法中的随机森林模型中的混淆矩阵	37
图 4-2 “强棒星系”和“非强棒星系”，下采样方法中的随机森林模型得到的参数重要性排序	38
图 4-3 “强棒星系”和“非强棒星系”，上采样方法中的随机森林模型中的混淆矩阵	38
图 4-4 “强棒星系”和“非强棒星系”，上采样方法中的随机森林模型得到的参数重要性排序	39
图 4-5 “强棒星系”和“非强棒星系”的 b/a、SERSIC 指数、g-r 颜色的分布直方图	40
图 4-6 “有棒星系”和“无棒星系”中的随机森林模型中的混淆矩阵	41
图 4-7 “有棒星系”和“无棒星系”中的随机森林模型得到的参数重要性排序	41
图 4-8 “有棒星系”和“无棒星系”的 b/a、SERSIC 指数、 r/r_{180} 的分布直方图	43

表目录

表 2-1 DESI Legacy Imaging Surveys 的三个测光巡天项目具体参数	9
表 2-2 本文所用样本	19

第1章 引言

星系是宇宙的基本组成单位，上可承接宇宙大尺度结构，下可细化为恒星、气体、尘埃、暗物质等当前热门方向，是天文领域中极其重要的一环，吸引了众多的科学研究。

1.1 星系的形态及基础物理参数

星系是由恒星，气体，尘埃和暗物质等组成的引力束缚系统，是天文学研究的基础。星系的形态是星系最直接的观测特征，是天文学家对星系的外观进行的一种视觉分类。星系形态有目视分类、模型化分类、非模型化分类和双色图分类等分类方法。最著名的一种星系形态分类是哈勃目视分类，是美国著名天文学家哈勃 (Edwin Hubble) 在 1926 年提出。哈勃将这些不同形状的星系排列在一条序列之上，因该序列形如音叉，又被称为哈勃音叉图 (参见图 1-2)。在哈勃音叉图中，不同形状的星系从左到右排列，分别对应椭圆星系 (E)、透镜星系 (S0)、旋涡星系 (S) 和不规则星系 (Irr) 四种。从形态学上看，椭圆星系呈椭球状，无明显的子结构特征，并可按视形状自圆而扁依次被分为 E0、E1……直至 E7 共八个次型。透镜星系是介于椭圆星系和旋涡星系之间的过度型星系，其作为一种无旋臂的盘星系，因其外形像侧视的透镜而得名。根据透镜星系中核心处是否有棒状结构，相应的分为 S0 或 SB0。旋涡星系有较明显圆盘形结构和旋臂等特征，中心处存在核球部分，部分旋涡星系还有棒，环等结构。首先，旋涡星系可以按照是否存在棒特征分为有棒旋涡星系 (棒旋星系) 和无棒旋涡星系两大类，分别对应哈勃音叉图中旋涡星系的两个分支。棒旋星系是本职工作所研究的重点，我们在图 1-1 给出了一个无棒旋涡星系 NGC 1300 (图 1-1 左) 和一个棒旋星系 NGC 1232 (图 1-1 右) 的图像，可以看到 NGC 1300 拥有的贯穿星系中心的显著的棒结构。此外，在哈勃音叉图中，旋涡星系 (包括有棒和无棒) 根据其旋臂的松紧程度和核球的大小，又被分为 a、b、c/d 等之类，其中 Sa 星系的旋臂最紧，核球最为明显，Sb 星系的旋臂较为松弛，核球相对较小，Sc/d 星系的旋臂最为松弛，核球最小。除了以上规则形态的星系外，其余外形不规则的星系则统归在不规则星系之中。

在星系形态学研究的早期，哈勃将位于音叉图左边的椭圆星系和透镜星系合并称为“早型星系” (Early-type galaxy)，将音叉图右边的旋涡星系和不规则星系称为“晚型星系” (Late-type galaxy)。之所以这样归类，是因为哈勃在最初认为这些不同的星系之间可能存在着某种演化关系，即椭圆星系最先诞生，随着演化转变成旋涡星系，最终演化为不规则星系。在今天，这一假设虽已被证明是错误的，即星系的演化并不是按照从椭圆星系到旋涡星系的顺序，但“早型星系”和“晚型星系”这一概念还是被保留了下来，沿用至今。在我们观测到的星系中，

大部分的亮星系是哈勃分类里的椭圆星系和旋涡星系，而光度较小的矮星系则多呈不规则形状。除矮星系之外，部分不规则星系的形状由星系的相互作用过程所引起，如著名的天线星系 (NGC4038 和 NGC4039) 就是由两个星系的并合过程所引起。



图 1-1 左图为 NGC 1300，右图为 NGC 1232

Figure 1-1 The left panel is NGC 1300 and the right panel is NGC 1232

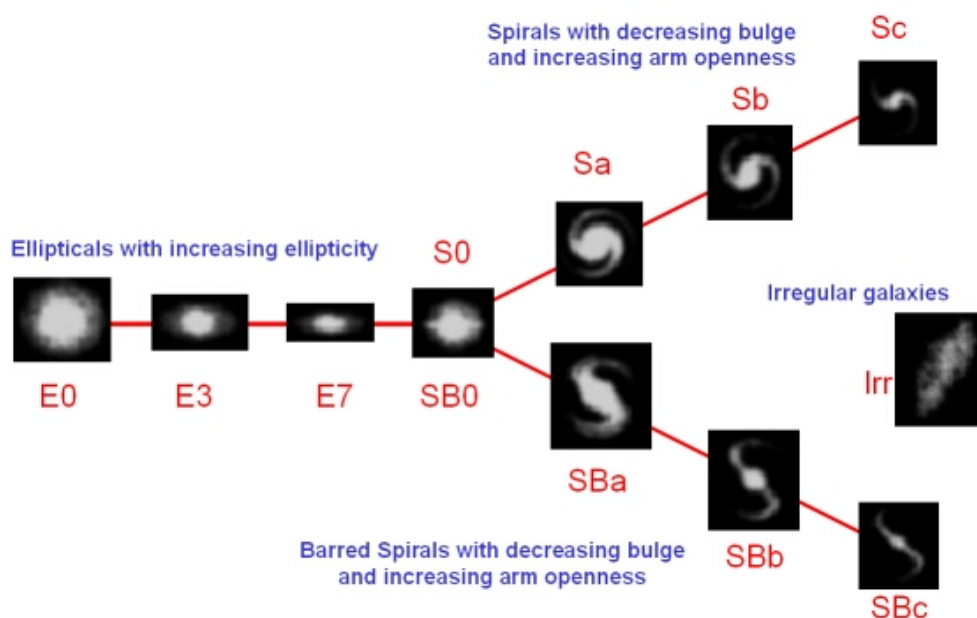


图 1-2 哈勃序列，也被称作哈勃音叉图 (来自 COSMOS)

Figure 1-2 Hubble classification, often referred to as the “tuning fork”

在星系的形态之外，天文学家也通过具体的物理参数来描述星系的性质。这些物理参数可以大致分为两部分：一部分是用于描述星系自身的性质，例如星系

的恒星质量 (M_*), 星系的物理尺度 (标高 scale height, 标长 scale length), 星系的恒星形成率 (Star Formation Rate, SFR), 星系的气相金属丰度 ($12+\log(O/H)$), 星系的表面亮度 (Surface Brightness), 星系的颜色 (例如 g-r color) 等。另一部分是星系所处环境的特征, 包括星系所属星系团/群的暗晕质量 (Halo Mass), 星系所属星系团/群的位置 (Central/Satellite), 第 N 近邻星系的投影距离 (Projected Distance of the Nth galaxy) 等用来描述星系周围密度的相关量。

在星系的内部物理特征中, 恒星质量和恒星形成率是描述自身恒星形成过程的最重要的两个物理量。一般来说, 星系恒星质量的测量可通过星系的恒星光度再结合质光比得到, 而质光比可以通过光谱能量分布拟合 (Spectral Energy Distribution Fitting, SED) 得到。恒星形成率作为单位时间内形成的恒星的质量, 可通过表征大质量恒星的多波段测光或光谱特征获得, 比如表征大质量恒星辐射的紫外连续谱和表征 HII 区电离气体的 $H\alpha$ 发射线, 都可以通过标定后估算星系 SFR 的值。

需要注意的是, 描述星系的内外两种物理参数之间并不是完全独立的, 其彼此存在着关联性。例如 Boselli et al. (2022) 的研究表明, 在星系的众多环境效应中, 星系团对星系的冲压剥离作用会将星系中的气体吹走, 进而降低恒星形成活动。Pettersson et al. (2023) 通过分析星系间的并合 (Merger) 现象, 发现并合会导致星系内的气体向中心聚集, 触发短时间的星爆 (Starburst) 现象, 在星爆阶段之后气体的不足会降低恒星形成活动, 进而使星系转入“熄灭” (Quench) 阶段。星系也有可能通过其它的物理过程进入熄灭阶段, 例如活动星系核 (Active Galactic Nuclei, AGN) 和超新星 (supernova) 的反馈作用, 这些过程会产生强烈的星风, 加热其周围的星际介质或吹走气体, 使气体无法聚集塌缩形成恒星。由于没有新形成的恒星贡献光度, 在天文观测中该类星系所发出的颜色较为偏红。Peng et al. (2010) 的研究将“熄灭”机制细化为质量“熄灭” (mass quenched)、并合“熄灭” (merger quenched) 和环境“熄灭” (environment quenched) 三种物理机制, 并给出了不同恒星质量下的“熄灭”星系, 所经历的“熄灭”是由哪部分“熄灭”机制占主导, 具体可见图 1-3。在小质量端部分是由并合“熄灭”和环境“熄灭”占主导, 随着质量的增大, 质量“熄灭”逐渐占据主导。

随着天文观测技术的进步, 天文学家发现哈勃序列中星系的形态分类, 与星系的物理性质也存在一定的相关性。像哈勃序列左侧的椭圆星系相对来说, 恒星质量较大, 比较年老, 颜色偏红, 气体较少, 星系中恒星形成较少, 星系的“熄灭”比例较多, 且多处于高密度区域; 而哈勃序列右侧的旋涡星系则相对质量较小, 年龄较年轻, 颜色偏蓝, 富含气体和尘埃, 星系中恒星形成较活跃, 且多处于低密度区域。

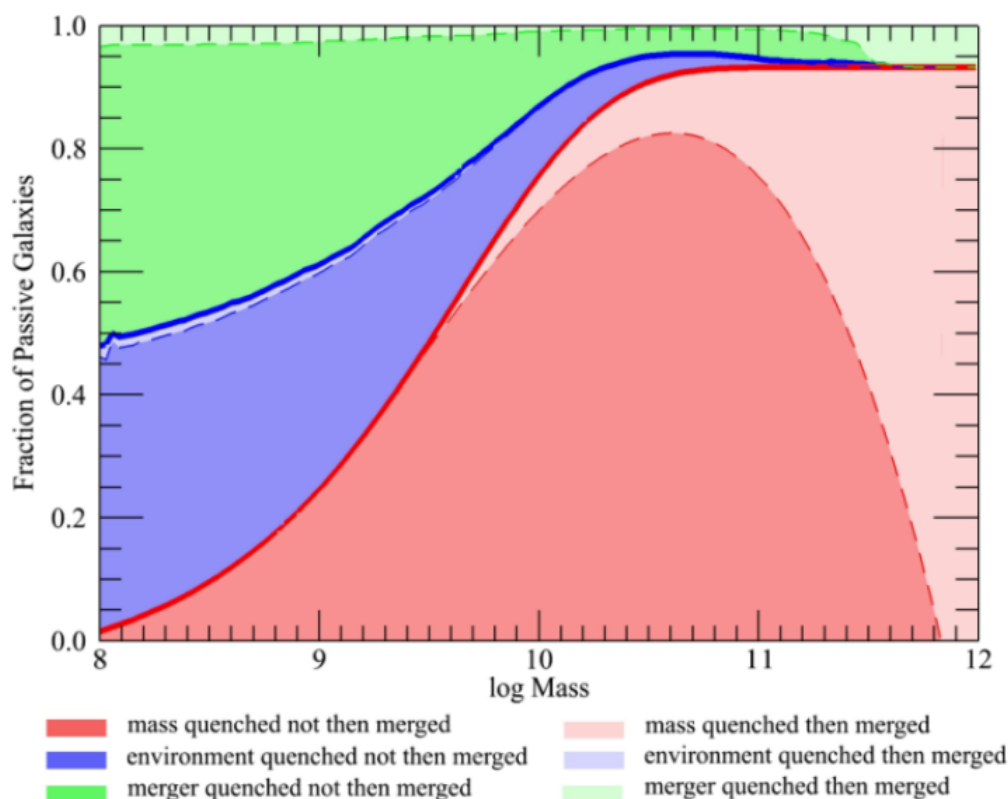


图 1-3 三种不同的“熄灭”机制，图源 Peng et al. (2010)
 Figure 1-3 Three different quench mechanism, from Peng et al. (2010)

1.2 棒旋星系

棒旋星系在宇宙中很常见，天文观测中得到的棒旋星系在总体旋涡星系中的比例从 30% 到 60% 不等，这取决于观测样本的红移和所观测的波长范围 (Nair et al., 2010a,b; Lee et al., 2012; Guo et al., 2023)，例如在近域宇宙 (Local Universe) 中可以达到 2/3 的比例 (Erwin, 2018)。不同的数值模拟工作也证明了天文观测得到的结论 (Rosas-Guevara et al., 2022; Zana et al., 2022; Bi et al., 2022)。

前人关于棒的研究，通常会先定义长度或强弱。棒旋星系中棒的长度不尽相同，有些可以和宿主星系的星系盘大小相比较，有些则相对小很多。在近域宇宙中，棒旋星系中的棒的长度分布在从不到 1kpc 到 15kpc 的范围内，中位值分布在 3-4kpc 左右 (Font et al., 2017)。造成棒长度差异的原因，可能是星系的质量分布，盘内恒星和气体的速度弥散，星系所处状态等因素的综合作用。故在工作中通常会根据棒的相对长度或宽度 (相对于宿主星系) 来定义棒的强弱，将棒旋星系分为强棒星系和弱棒星系 (Géron et al., 2021)。强棒和弱棒星系之间也存在一定差异，Knapen et al. (1995) 认为强棒星系中的棒具有较低的 SFR，且更集中在棒的两端，而弱棒则相对的增强恒星形成。

当前认为，棒旋星系中的棒会将星系中的气体输送到星系中心区域 (Lin et al., 2017)，增强恒星形成，促使中央大质量黑洞增大，进而增强 AGN 活动，同时也会对星系盘上的角动量和能量进行再分配，推动星系盘与暗物质晕的相

互作用 (Athanassoula, 2003; Kormendy et al., 2004)。另外, 棒还可能与盒型/花生 (boxy/peanut) 形状或 X 形状 (X-shape) 的伪核球 (pseudo-bulge) 的形成有关 (Bureau et al., 1999; Merrifield et al., 1999; Erwin et al., 2017), 是星系长期演化 (secular evolution) 的一个主要催动力 (Sellwood, 2014)。

目前关于棒的物理机制, 比较得到认可的一种是星系中的棒使气体内流, 加剧中心区域短时间的恒星形成, 但在此之后, 星系由于缺乏气体, 会进入到长时间的恒星形成较低的阶段, 也就是“熄灭”阶段, 在颜色上会表现得较为偏红 (Kim et al., 2017; Consolandi et al., 2017)。在更细致的研究中, Knapen et al. (1995) 发现在 K 波段更容易观测到棒, 佐证了本地棒旋星系中的棒通常为年老星族这一观点。Kruk et al. (2018) 发现有棒星系的盘比无棒星系的盘要偏红, 且有棒星系的棒区域要比自身的盘区域更红。也有研究得到了在大质量, 颜色偏红且气体含量较低的星系中, 棒出现的频率较高的结论 (Masters et al., 2012; Cervantes Sodi, 2017)。

棒的起源现在仍不明确, 其形成机制较为复杂。有关数值模拟的工作发现, 棒既可以在恒星盘中由引力不稳定性自发形成 (Tahmasebzadeh et al., 2022), 又可以由近距离的飞掠 (Flyby) 和并合触发形成 (Martinez-Valpuesta et al., 2017)。同时, 如 1.1 节所提到星系的内部和外部两种物理参数之间存在关联, 同样有很多研究关注棒旋星系的环境与棒之间的联系。在数值模拟中, 小尺度的环境可能对星系中棒的演化产生重要影响。比如, 星系对 (galaxies pairs) 中产生的强相互作用可能会摧毁棒结构 (Tawfeek et al., 2022); 星系-星系之间的飞掠 (Zana et al., 2018) 或并合 (Ghosh et al., 2021) 现象可能会使棒的强度减弱, 长度缩短, 甚至将棒摧毁。也有数值模拟工作指出星系团潮汐力可以在星系团中心处触发棒的形成, 而星系团外围则无法触发, 因此在星系团中心可以发现更多的棒星系 (Łokas et al., 2016; Łokas, 2020)。

在观测上, 棒与星系所处的环境之间是否存在相关性目前尚无明显的共识。较多工作认为棒与环境并无明显相关性或并未起到重要作用, 例如 van den Bergh (2002) 通过研究棒旋星系在星系群 (Group), 星系团 (Cluster) 和星系场 (Field) 内的比例, 没有发现不同的环境对棒的形成有较大差别; Barazza et al. (2009) 发现星系团和星系场的有棒星系的比例, 以及棒的长度和宽度都很相似; Aguerri et al. (2009) 通过研究斯隆数字巡天第五期 (Sloan Digital Sky Survey Data Release 5, SDSS DR5) 的样本, 发现有棒星系和无棒星系所处区域的局域密度没有差别, 且棒的长度和强度也都与局域密度无关; Marinova et al. (2009) 在对红移为 0.165 的阿贝尔 901/2 超星系团 (Abell 901/2 supercluster) 的工作中, 也发现星系团的环境并不显著影响有棒星系的比例; 因此, Sarkar et al. (2021) 认为棒的形成更多的受星系内部过程的影响, 而不是环境。另一方面, Tawfeek et al. (2022) 研究了宇宙大尺度下的星系团的结构与棒的相关性, 发现在星系团外围的旋涡星系或星系团位力半径内的椭圆星系中, 含棒的星系比例最多。

造成以上研究结论不统一的重要原因之一是星系中棒特征的证认目前仍然

依赖于人眼分类。这样的人类分类判据一方面限制了有棒星系样本的大小（参见 2.1 节），另外一方面也可能使得不同的有棒星系样本中存在不同的人眼选择判断偏差（参见第 4 章）。

1.3 本文的研究目的

正如 1.1 节中所述，当前对于星系中的恒星形成过程与所处环境之间的关系已得到充分研究。然而星系中的棒特征与环境之间的关系仍尚不清楚。如果棒的形成和演化过程也与星系所处的环境密切相关，那么就需要重新审视观测中棒和星系整体恒星形成性质之间的关联。目前，还缺少将星系中的棒特征与 SFR、环境之间的因果关系进行分离的详尽研究。同时由于星系中弱棒特征的判断较为困难，较多研究都只集中在有棒星系和无棒星系的差别中，对于棒星系（强棒和弱棒）更细致的分析还有很大的研究空间。

如今，大规模的公众科学计划提供了更大的旋涡星系样本（参见 2.1 节），给了我们重新审视旋涡星系形态（如棒特征）、恒星形成以及环境的机会，以往微弱或不曾发现的物理，都有可能随着更多的样本和更细致的样本，慢慢展现出来。此外，众多的基于多波段测光数据提供的星系物理参数测量的增值星表（参见第 2 章），也大大减轻了我们进行系统对比分析的工作量。

在恒星形成特征以外，我们同样感兴趣于更多的星系物理参数（特别是观测参数）与棒之间的关系。倘若棒与星系其它的物理参数存在良好而复杂的相关关系，那么一方面我们可以根据星系中不依赖于人眼的物理参数的测量，来有效的预言星系中棒这个深度依赖于人眼的形态特征，另外一方面，我们可以通过棒特征和这些物理参数的内在关联进一步深入分析棒在星系的演化过程中所扮演的角色。

本文的研究目的正是利用公众科学计划提供的大样本棒旋星系，开展棒、SFR 以及环境三者之间关联的详细统计分析。此外，通过现代的机器学习算法（随机森林分类，3.1.2 节）来深入分析星系的物理参数与棒的相关性。在下一章中我们将介绍本文所用数据集。

第 2 章 本文使用的数据集及样本

本文使用的数据来自以下四个数据集：Galaxy Zoo DECaLS (GZD) 数据集、GALEX-SDSS-WISE Legacy Catalogue-X2 (GSWLC-X2) 数据集、星系群 (Yang's Group catalog) 数据集和 NSA (NASA-Sloan Atlas) 数据集，下面将详细介绍各数据集以及本文中所用的样本。

2.1 Galaxy Zoo DECaLS

2.1.1 Galaxy Zoo

星系的形态类别是了解星系的形成和演化的重要工具。在最初的科研工作中，星系样本数量较少，方便天文学家进行逐个细致的目视检查。然而随着新的巡天项目逐渐发展，海量天文数据的涌来不仅极大的推动了科研的进展，也逐渐成为了科学家们幸福的“烦恼”。例如始于 2000 年的斯隆数字巡天项目 (Sloan Digital Sky Survey, SDSS)，同时具备多波段测光和多目标光纤光谱观测的能力，作为天文学史上最具有影响力的巡天项目之一，提供给天文学界海量的天文图像和天文光谱。面对前所未有的天文图像数量，以往科学家所用的逐个目视方法也变得不再可能，广泛收集公共投票的“星系动物园”项目 (Galaxy Zoo project, GZ project) 应运而生。Galaxy Zoo 项目于 2007 年 7 月开始，由英国牛津大学的天文学家发起，并得到了全球数十万志愿者的支持，收集了超过 4 千万的形态学投票，极大的推进了科研的脚步。该项目不再仅面对专业的天文学家，而是将目光投向了公共志愿者。只要经过平台一定的训练之后，就可以参与到项目中发布的形态学投票。在公共志愿者开始参与时，平台会针对单张图片，给出一系列的形态学问题，如图 2-1。在志愿者给出不同的回答时，平台会根据事先设置好的决策树流程，提供下一个问题引导志愿者进行回答，以此类推，直至完成该张照片的所有分类问题为止。协助 SDSS 巡天还只是公共科学家助力科研的开始，Galaxy Zoo 平台在后续也极大帮助了 HST, DECaLS 等众多项目，不仅提高了公众对天文学的兴趣和参与度，也为科学研究提供了宝贵的数据和发现。

Galaxy Zoo 项目的出现，带来的一个显而易见的好处是节省了大量的时间成本，另一方面，如果没有对同一个星系多次独立的形态分类，就很难确定分类有多大的可信度，Galaxy Zoo 项目通过对样本中的每个星系进行大量的独立分类，可以对误差进行量化，也是使其成为了研究星系演化的强大工具的一个重要原因。

Galaxy Zoo 平台为不同的望远镜、巡天项目和数值模拟项目提供了不同的决策树流程，包括 HST, SDSS, CANDELS, UKIDSS, Ferengi, DECaLS, illustris, GAMA-KiDS 等。具体项目的决策树流程详见 [galaxy zoo.org/gz-trees](http://galaxy.zoo.org/gz-trees)。

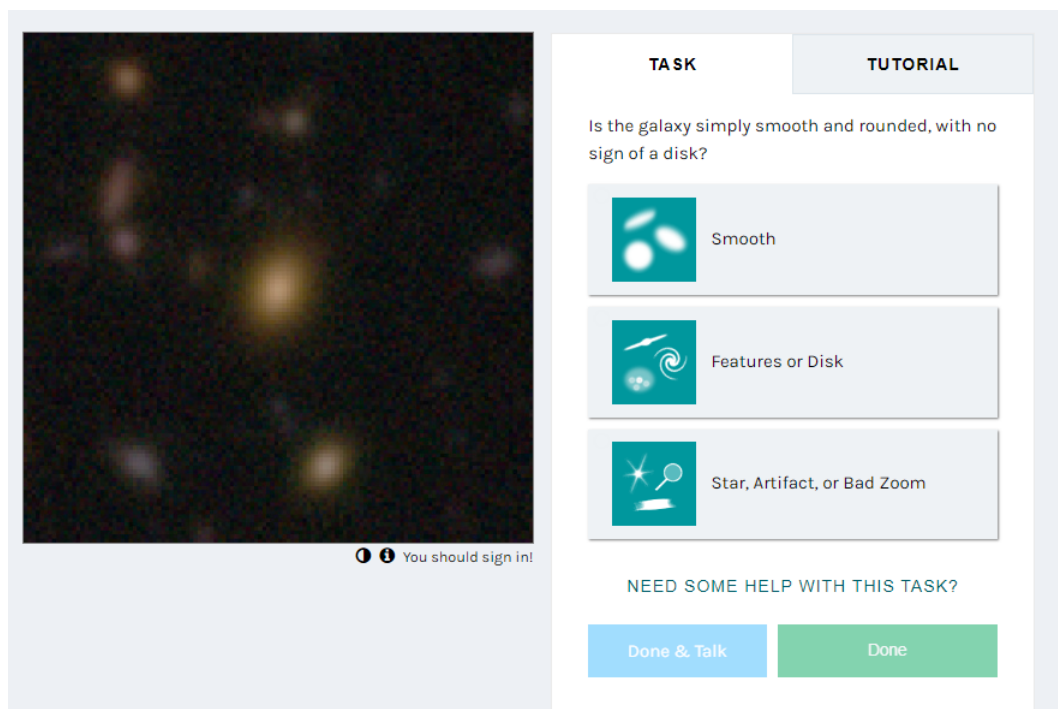


图 2-1 Galaxy Zoo 网站中的分类投票页面

Figure 2-1 The classification voting page on the Galaxy Zoo website

2.1.2 DESI legacy survey

DESI (Dark Energy Spectroscopic Instrument) 巡天作为一个具有重要意义的大型国际合作项目，旨在通过收集千万级星系发出的光谱，绘制宇宙三维“地图”，从而揭开暗能量之谜。

而在 DESI 光谱巡天项目启动之前，为了选取进行光谱观测的目标源，暗能量光谱遗珍成像巡天 (DESI Legacy Imaging Surveys, Dey et al. (2019)) 对 DESI 天区进行了测光巡天。其使用位于 Kitt Peak 国家天文台和 Cerro Tololo 美洲天文台的望远镜，对超过 14000 平方度的高银纬天空进行成像，得到了三个光学波段 (g, r, z) 的测光数据。其测光天区覆盖范围限制在了赤纬-18 度至 +84 度，银纬高于 +15 度或小于-15 度区域。

DESI 测光巡天项目由三个不同的望远镜上的成像项目组成，分别是北京-亚利桑那天文巡天 (The Beijing-Arizona Sky Survey, BASS)，暗能量相机遗珍巡天 (The Dark Energy Camera Legacy Survey, DECaLS) 和 Mayall z 波段遗珍巡天 (The Mayall z-band Legacy Survey, MzLS)。此三个测光巡天项目各自覆盖天区可见图 2-2¹。三个项目的基本参数可见表 2-1。

其中，BASS 是由中国科学院国家天文台和美国亚利桑那大学斯图尔德天文台联合开展的一项巡天项目，使用了位于 Kitt Peak 国家天文台的 Bok 2.3 米望远镜，通过 g 和 r 两个波段对覆盖了北银冠 (Northern Galactic Cap) 区域约 5400 平方度的区域进行成像。该巡天于 2015 年 12 月发布了早期数据 (EDR)，分别于

¹<https://www.desi.lbl.gov/imaging-surveys/>

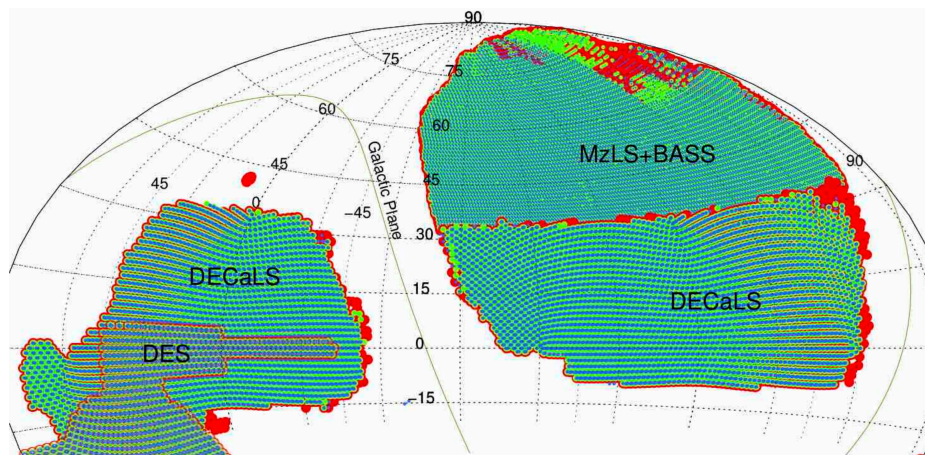


图 2-2 BASS、DECaLS 和 MzLS 三个测光巡天项目天区覆盖图

Figure 2-2 Sky coverage map of the BASS, DECaLS, and MzLS photometric survey projects.

2017 年 1 月、2017 年 12 月和 2019 年 8 月发布了第一批、第二批和第三批数据。

MzLS 由美国国家光学天文台负责实施，利用同样位于 Kitt Peak 国家天文台的 Mayall 4 米望远镜开展 z 波段的观测。该测光巡天与 BASS 覆盖相同的天区范围但提供了不同的波段观测数据，于 2016 年 2 月开始正式观测，至 2018 年 2 月结束。值得注意的是，该巡天所用的 Mayall 4 米望远镜，在进行改造后也是后续 DESI 巡天中所用望远镜。

DECaLS 是美国劳伦斯伯克利国家实验室负责实施，利用智利 Cerro Tololo 美洲天文台的 Blanco 4 米望远镜上的暗能量相机 (Dark Energy Camera, DECam (Flaugher et al., 2015)) 进行三个波段 (g, r, z) 的观测，于 2019 年 3 月完成了所有的观测。暗能量相机具有大约六边形的 3.2 平方度的视场，每个像素覆盖 0.262 个角秒，g、r、z 三个波段的中位值点扩散函数的半高全宽 (FWHM) 分别为 1."29、1."18 和 1."11。且 DECaLS 项目覆盖了 10480 平方度的天区，占整个 DESI 测光天区 14000 平方度的将近 75%，是其最重要的组成部分。DECaLS 项目在 400-1000nm 范围内具有大视场和高灵敏度的优点，能拍出比 SDSS 更深，分辨率更高的星系图像。

表 2-1 DESI Legacy Imaging Surveys 的三个测光巡天项目具体参数

Table 2-1 The three photometric survey projects of DESI Legacy Imaging Surveys

	DECaLS	BASS	MzLS
望远镜/相机设备	Blanco/DECam	Bok/90prime	Mayall/Mosaic-3
观测波段	g, r, z	g, r	z
天区	NGC($\delta \leq +32^\circ$) + SGC	NGC ($\delta \geq +32^\circ$)	NGC ($\delta \geq +32^\circ$)
天区覆盖面积	约 10480 平方度	约 5400 平方度	约 5400 平方度
分辨率	0."262	0."454	0."262

2.1.3 Galaxy Zoo DECaLS

Galaxy Zoo DECaLS(GZD) 项目是 Galaxy Zoo 团队根据 DECaLS 的测光图像数据开展的新一代的星系形态分类的公众科学项目。具体来说, DECaLS 项目组将 DECaLS DR1、DECaLS DR2 和 DECaLS DR5 三期项目的测光图像, 放于 Galaxy Zoo 平台合作并推出共三期的 GZD 项目, 分别是 GZD-1, GZD-2 和 GZD-5 项目。GZD-1 项目从 2015 年 9 月开始到 2016 年 2 月结束, 每个源普遍收到了 40 个至 60 个投票, GZD-2 项目从 2016 年 4 月开始到 2017 年 2 月结束, 每个源平均收到了 40 个左右的投票。其中 GZD-1 和 GZD-2 共用相同的决策树, 共对 99109 个星系进行了形态学分类。GZD-5 项目从 2017 年 3 月开始到 2020 年 10 月结束, 其应用了与 GZD-1 和 GZD-2 不同的新的决策树, 更加针对棒和旋臂等微弱特征 (Walmsley et al., 2022)。由于 GZD-5 的数据样本远大于 GZD-1 和 GZD-2, 在 2019 年 6 月前的分类中, 每个 GZD-5 的源都近似收到了 40 个投票。但在 2019 年 6 月后, 为减少时间和人力成本, 其工作团队通过一种主动学习算法, 优先选取“信息量大”的星系交予 Galaxy Zoo 平台的志愿者进行分类, 该类星系仍然收到 40 个左右的投票数, 而剩余的源则只需最少接收 5 个分类投票即可。该团队后续将 GZD-5 中志愿者的投票用来训练一个贝叶斯卷积神经网络模型, 而“信息量大”的星系是指在训练模型中, 最能提升模型性能的那些星系。

在三期 GZD 项目中, 形态学投票所依据的决策树的建立和优化是关键。在 GZD-1 和 GZD-2 中, 决策树(如图 2-3 所示)所给出的问题首先聚焦于所给出的源是椭圆星系、旋涡星系或恒星/干扰源。排除掉恒星/干扰源后, 决策树会根据志愿者选择的答案, 进一步给出不同的问题。例如当志愿者选择椭圆星系后, 决策树会询问该源具体有多圆(完全椭圆、中间值、雪茄状三个选项), 并接下来询问该源有无并合现象或潮汐尾等特征, 以及最后询问有无类如环、尘埃带等特殊现象; 而旋涡星系这一分支则对应更多更细节的问题, 例如决策树会首先询问旋涡星系的盘是正向于观测者还是侧向于观测者, 若侧向的话是否能观察到核球的存在以及核球的形状, 若正向时则会询问志愿者有无棒的特征、有无旋臂的特征、旋臂松紧程度、旋臂的数量、核球大小的占比等等问题, 最后的问题与椭圆星系类似, 都会询问志愿者该旋涡星系有无并合现象或潮汐尾等特征, 以及有无类如环、尘埃带等特殊现象。

而 GZD-5 所应用的决策树, 则在棒、核球以及星系并合等方面的问题上进行了优化, 能够更灵敏的提取相关的信息。在刚开始面向志愿者的问题中, GZD-5 与之前工作一样, 在询问椭圆星系/旋涡星系后, 若志愿者回答旋涡星系, 则会继续提问该星系的盘是正向还是侧向于观察者。但在关于棒特征的问答上, GZD-5 的决策树给出了更为具体的选择, 将原本的“是、否”的选择转化为“无棒、弱棒、强棒”的选择, 将原本的二分类问题细化成了更具统计分析意义的三分类问题, 更便于人们对棒进行细致的研究。同时在后续询问核球的信息上, 志愿者可选择的答案也从简单的“不存在、存在且明显、存在且主导”的三分类变成了“无核球、小核球、中等核球、大核球、核球主导”的五分类问题。以及关于并

合现象的回答，除了原有的“有并合现象、无并合现象”的答案可以选择外，志愿者还可以选择“无并合但有主扰动、无并合但有次扰动”。更具体更细致的选项可以使团队获得更具有信息量的投票，有利于后续关于棒、核球、并合等特征的分析。在三期项目中，公共志愿者总共对这些 313789 个同样存在于 SDSS 主星系样本 (定义为 $r < 17.77$ 等) 中的星系进行了形态学分类的投票，收集到的投票总数达到了惊人的 749 万 6 千个。其中关于星系图像上的棒，旋臂，核球等子结构的存在与否的投票，以及对存在的结构的强弱，数量等问题进行的回答，都是该 GZD 项目在星系形态学分析大展身手的的原因。

在卷积神经网络模型训练完成后，该团队应用此模型重新预测了 GZD 三个项目中所有的星系在面对 GZD-5 决策树不同节点时的答案。本文所应用的 GZD 数据集即为该团队应用深度学习所预测的数据集。

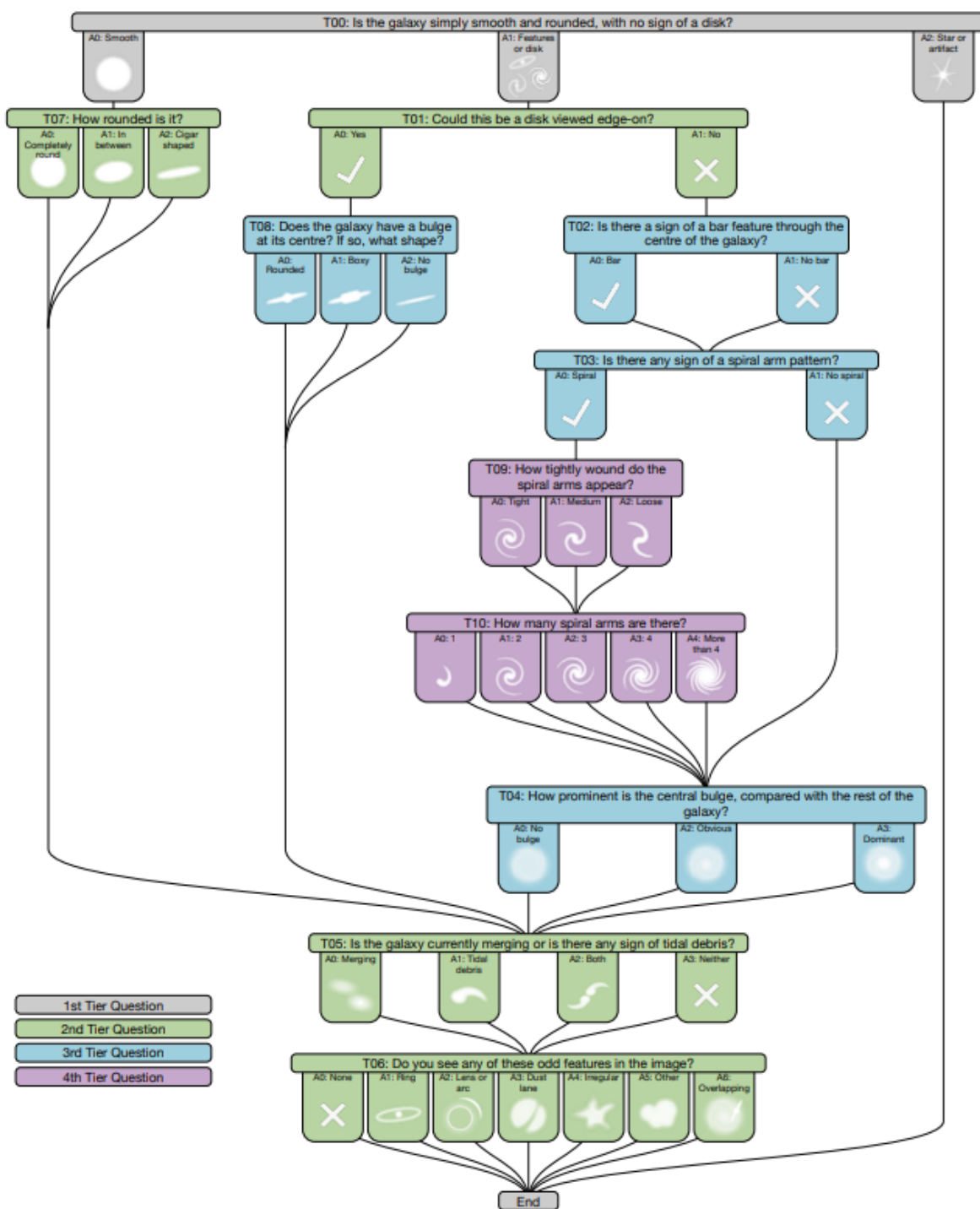


图 2-3 GZD-1 和 GZD-2 所用的决策树
 Figure 2-3 Decision trees used for GZD-1 and GZD-2 campaigns

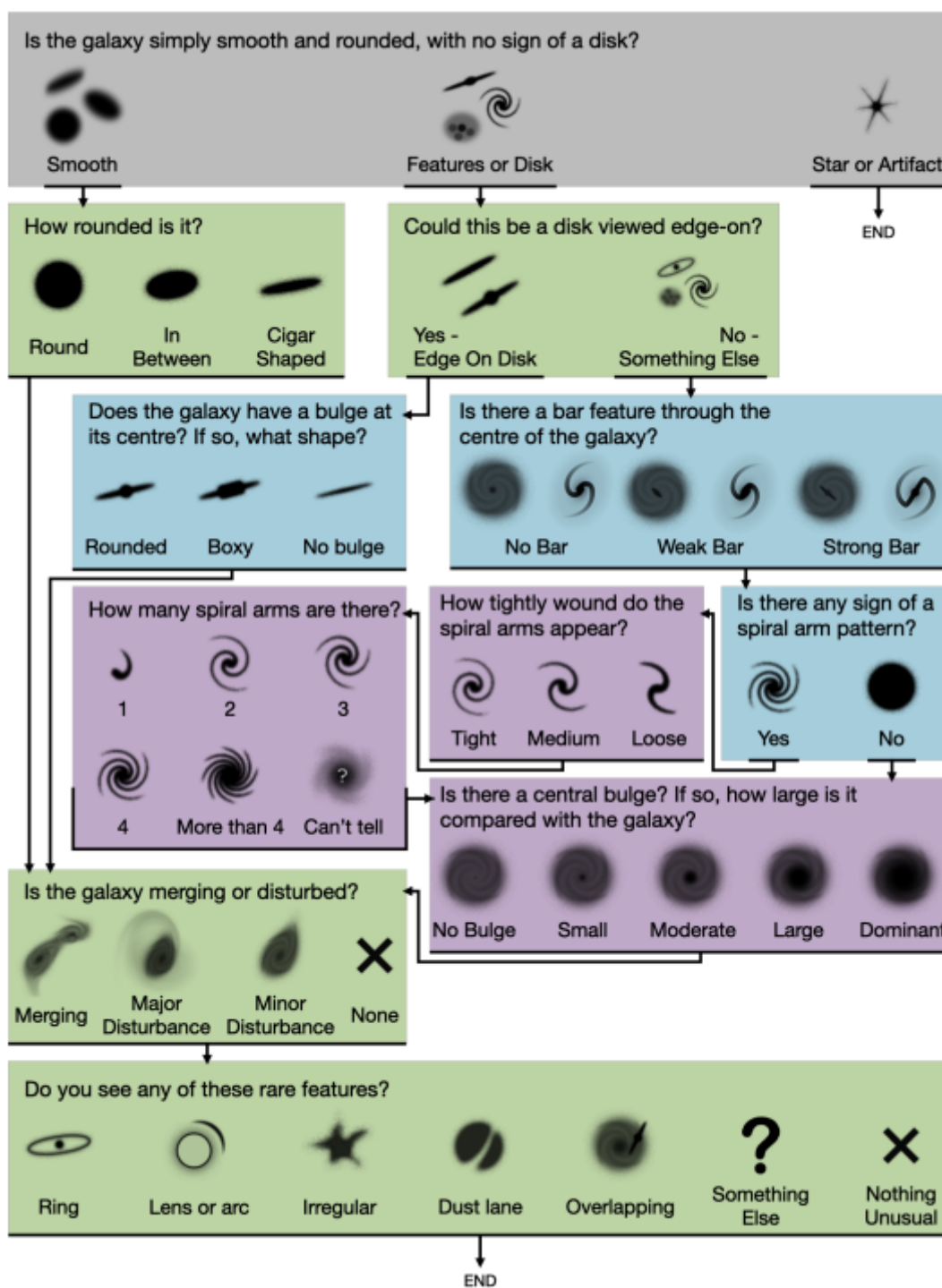


图 2-4 GZD-5 所用的决策树
Figure 2-4 Decision tree used for GZD-5 campaigns

2.2 NSA 数据集

NSA 数据集²是一个包含本地星系图像和参数的目录，这些图像和参数来自 SDSS 巡天，同时添加了星系演化探测器 (Galaxy Evolution Explorer, GALEX) 带来的紫外波段部分的光谱能量分布 (Spectral Energy Distribution, SED) 的数据。NSA 样本包含了五个光学波段 (u、g、r、i、z) 和两个紫外波段 (NUV、FUV) 的测光数据。其中光学波段数据来自 SDSS 巡天，紫外波段数据来自 GALEX 太空望远镜。

在 NSA 数据集中，提供了星系的红移 (redshift) 和位置信息 (RA, DEC); SDSS 巡天中独有的星系编号信息 (RUN-CAMCOL-FIELD); 星系的亮度信息，例如七个不同波段 (u、g、r、i、z、NUV、FUV) 的经过 K 改正后的绝对星等 (Absolute Magnitude) 和 AB 星等; 星系的形态信息，如 Sersic 指数，其作为拟合星系面亮度轮廓的 Sersic 模型的参数，当 Sersic 指数 = 1 时，代表指数轮廓，当 Sersic 指数 = 4 时，代表 de Vaucouleurs 轮廓 (通常用来拟合椭圆星系的面亮度轮廓)。也就是说，Sersic 指数越大，代表星系的聚集度 (Concentration) 越高。星系的轴比 (b/a)，为星系的短轴方向的长度除以星系长轴方向的长度。星系的对称度 (Asymmetry) 信息，为 r 波段观测下星系的对称度，通过将星系图像旋转 180° 后相减得到; 星系的尺度信息，如 Petrosian 50 和 Petrosian 90，分别代表 r 波段下 50% 和 90% 的 Petrosian 光度的位置，可作为星系的某种半径信息; 星系的质光比 (mass-to-light ratio) 信息; 星系的 D4000 信息; H α 、H β 以及其它谱线的流量信息。

2.3 GSWLC-X2

GALEX-SDSS-WISE Legacy Catalogue 是一个包含约 700,000 个 SDSS 红移小于 0.3 且星等星系的物理性质目录 (Salim et al., 2016, 2018)。它使用 GALEX、SDSS 和 WISE 三部分的数据，通过紫外/光学 SED 来获取星系的物理性质 (例如恒星质量、尘埃减光和恒星形成率等)。

该目录有两个版本: GSWLC-1 和 GSWLC-2。两个版本都包含相同的源 (完全相同数量的对象按相同顺序排列)，并且它们都使用相同的测光信息。主要区别在于 GSWLC-1 中的恒星形成率分别来自紫外和光学的 SED 拟合 (无中红外约束) 和 WISE 22 微米光度测量。而 GSWLC-2 中的恒星形成率来自联合的紫外、光学和中红外 SED 拟合的更精确的 SFR。

同时根据紫外测光的深度，又可以将 GSWLC 分为三个子目录: GSWLC-A, GSWLC-M, GSWLC-D。A、M、D 三个子目录分别覆盖了 SDSS DR10 目标的 88%、49% 和 7%，如果同一星系被具有不同深度的紫外波段观测到，那么可能同时出现在不同的目录中。此外，命名为 GSWLC-X 的主目录采用了 A、M、D 三个目录中最深的那部分星系 (共 659229 个)，覆盖了 SDSS 主星系样本的 90%。

²<http://www.nsatlas.org/data>

2.4 星系群数据集

在 Yang et al. (2007) 关于斯隆数字巡天项目第七期 (SDSS DR7) 的星系群数据集³中, 应用了改进的基于暗物质晕模型的星系群搜索算法, 给出了每个星系所属的星系群, 也包括星系群的坐标和红移, 星系群的光度, 星系群的恒星质量和星系群所属暗晕的暗晕质量等信息。星系群搜索算法使用迭代的方式来估计每个星系群所处暗晕的质量、大小和速度弥散, 然后使用这些参数来确定该红移空间中的群成员的资格。具体来说, 该算法首先使用传统的 FOF 算法 (Friends of Friends) 对星系进行分配, 确定潜在的星系群中心, 并且估计中心处的特征光度。此处只计算 r 波段绝对星等亮于 -19.5 mag 的星系。然后该算法使用星系群的平均质光比对每个星系群分配一个临时质量 (tentative mass), 通过这个临时质量来估计其所处暗晕的大小、质量和速度弥散度, 并由此确定该红移空间内的成员星系, 每加入一个星系则需要重新迭代一次, 直至没有新的成员星系加入。对于该星系群所处暗晕的暗晕质量, 其算法可以给出两种计算方式得到的值, 一种是通过星系群内所有 r 波段绝对星等亮于 -19.5 mag 的星系的恒星质量计算得到, 另一种是通过星系群内所有 r 波段绝对星等亮于 -19.5 mag 的星系的恒星光度计算得到。

星系群搜索算法应用在了诸多的天文观测项目, 如 2dFGRS、SDSS DR4、SDSS DR7、2MASS 红移巡天等, 为识别和表征星系群提供了一种可靠的方法, 可将研究结果用于进一步研究星系群的特性及其在星系演化中的作用, 对于理解宇宙的大规模结构以及星系的形成和演化非常重要。

我们将 DECaLS 天区、GSWLC 天区和星系群数据集的覆盖天区以 Aitoff 投影的形式画出, 可见图 2-5。其中灰色图例为 GSWLC 数据集的覆盖天区, 蓝色图例为星系群数据集的覆盖天区, 绿色图例为 DECaLS 覆盖天区, 红色图例为三个数据集的重叠覆盖天区。由于 DECaLS 团队本身就使用 NSA 样本对 DECaLS 巡天中的目标源进行识别, 故图 2-5 中没有画出 NSA 样本的天区覆盖范围。

³<https://gax.sjtu.edu.cn/data/Group.html>

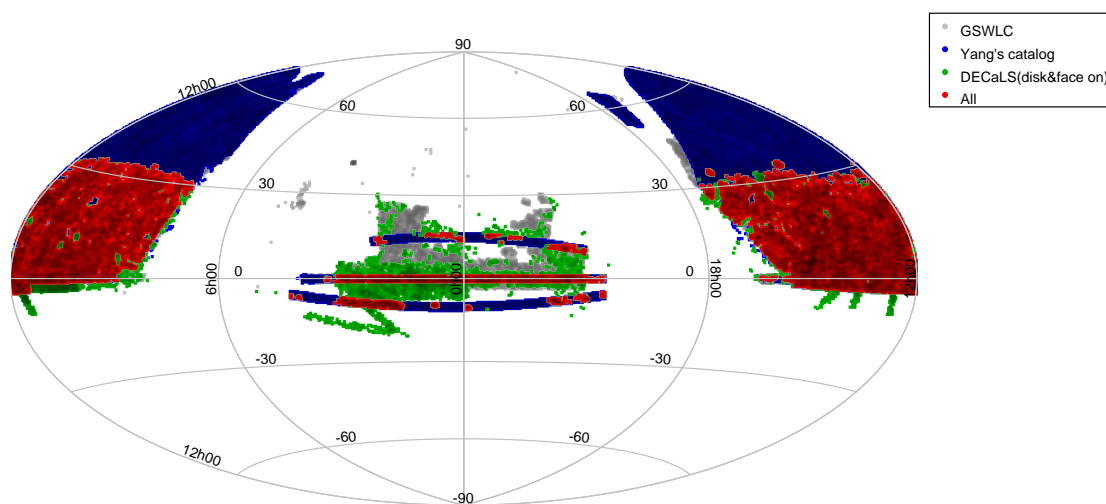


图 2-5 各数据集的 AITOFF 投影覆盖范围

Figure 2-5 The Aitoff projection of each data set used in this study

2.5 所用样本

本文研究的重点是星系中的棒结构，星系中的棒相关信息由 GZD 数据集提供。在 GZD 数据集中，先选取正面朝向的旋涡星系，然后再进行棒的特征判断，也就是再将其划分为强棒，弱棒和无棒星系来进行分析。对应到 GZD-5 的决策树中，我们选取“旋涡星系 (Features or Disk)”以及“正面朝向 (Not Edge On)”两个标签的投票比例相乘大于 50% 的样本，将其作为可分析棒星系的样本。该样本数目有 56912 个源，考虑到该样本提供棒这个形态学 (Morphology) 信息，方便起见，我们将此样本命名为“Morph”样本。基于“Morph”样本，本文进一步根据模型预测的投票数据进行分类，若模型对单个星系的“无棒”选项投票比例大于 50%，我们则认其为“无棒星系”（数目：33486）；若剩余的星系中，投票为“弱棒”的比例大于“强棒”的比例，则认为其是“弱棒星系”（数目：16703），反之则为“强棒星系”（数目：6723）。以上“无棒”、“弱棒”和“强棒”的样本选择流程图可参考图 2-6。

星系的恒星形成信息来自 GSWLC-X2 数据集，基于覆盖范围和 SFR 的精确度的原因，在工作中我们选取了 GSWLC-X2 这个子目录中的星系恒星质量和恒星形成率等信息。同时进一步计算得出星系的 $sSFR$ ($sSFR=SFR/M$)。我们将“Morph”样本与 GSWLC-X2 进行交叉匹配后。此步过后样本数目为 43221 个，为方便起见，将此命名为“Morph-SFR”样本。

星系的环境信息来自星系群样本，在其工作中根据三种不同组合的红移信息 (只有 SDSS 红移、SDSS 红移 + 2df 红移、SDSS 红移 + 2df 红移 + 最近邻红移插值 + ROSAT X-ray 星系团数据集中的红移) 以及两种不同的绝对星等 (petrosian 星等、model 星等)，给出了具有细微差别的 6 种不同的星系群样本。考虑到最近邻红移插值带来的不确定性以及 model 星等覆盖星系总体流量，我们将“Morph”样本与 SDSS 红移 + 2df 红移 + model 星等的子样本来进行交叉匹配，获得了

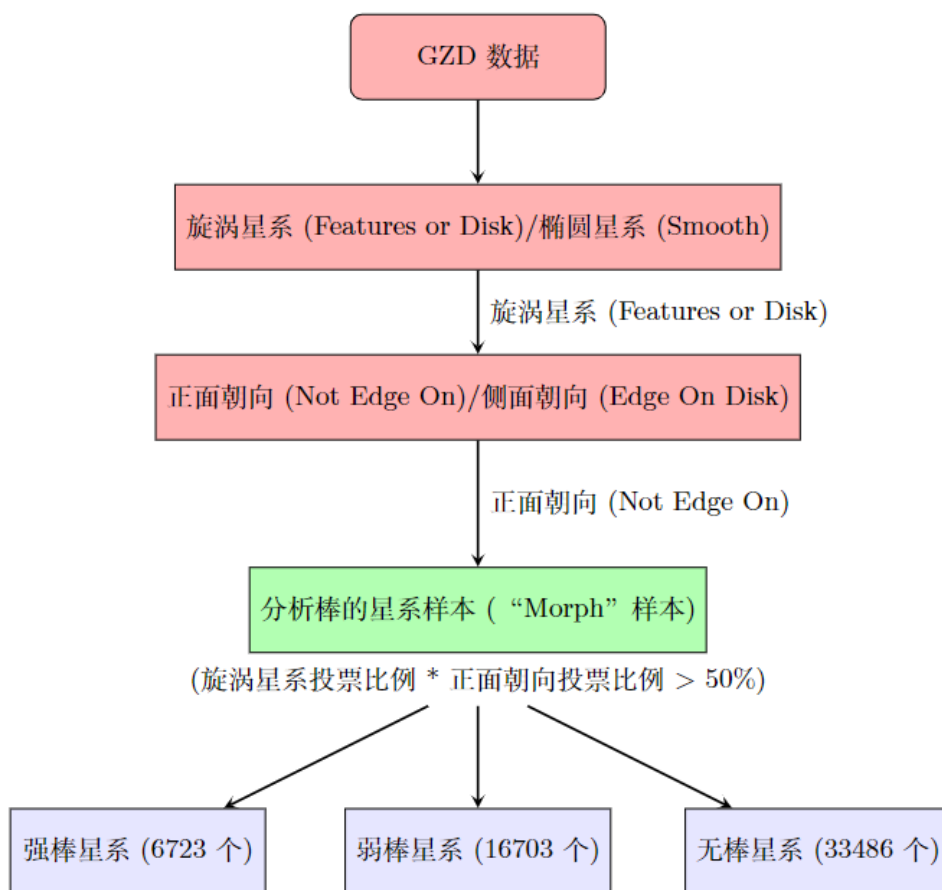


图 2-6 获取“Morph”样本所用流程图。其中判定为强棒星系需遵循“无棒投票比例 < 50% 且强棒投票比例 > 弱棒投票比例”，判定为弱棒星系需遵循“无棒投票比例 < 50% 且强棒投票比例 < 弱棒投票比例”，判定为无棒星系需遵循“无棒投票比例 > 50%”

Figure 2-6 The flow chart of getting the “Morph” sample. The galaxy which $f_{no\bar{b}} < 50\%$ and $f_{strong\bar{b}} > f_{weak\bar{b}}$ are labelled as “strong bar”. The galaxy which $f_{no\bar{b}} < 50\%$ and $f_{strong\bar{b}} < f_{weak\bar{b}}$ are labelled as “weak bar”. The galaxy which $f_{no\bar{b}} > 50\%$ are labelled as “no bar”

35807 个星系所处暗晕的信息，将此样本命名为“Morph- M_h ”样本。对于没有匹配到暗晕信息的源，主要原因是星系群样本中只计算 r 波段绝对星等亮于-19.5 等的星系的所属暗晕质量 (详细见 Yang et al. (2007))，另外还存在部分天区没有覆盖的情况。我们选取了星系群内恒星质量计算得到的数据作为该星系所处暗晕质量 (M_h)。通过暗晕质量 (M_h)，星系群红移等信息计算星系群中 180 倍宇宙平均密度处的半径 (r_{180})，视其为星系群的位力半径。具体公式为：

$$r_{180} = 1.26h^{-1}\text{Mpc} \left(\frac{M_h}{10^{14}h^{-1}M_{\odot}} \right)^{1/3} (1 + z_{group})^{-1} \quad (2-1)$$

其中 M_h 为星系群的暗晕质量， z_{group} 为星系群的红移值， h 为哈勃常数。为简单起见，本文的 h 取为 0.73。我们定义星系群中恒星质量最大的星系为星系群中心，计算其余星系与此中央星系的投影距离 r ，再将 r 用星系群的位力尺度 r_{180} 归一化后得到的 r/r_{180} 视为星系所处星系群里的位置。此外，我们通过星系群数据集中的星系的红移得到每个星系的视线方向退行速度 V_{los} ，其与星系群视线方向退行速度之差为 $|\Delta V|$ 。我们采用了 Beers et al. (1990) 提出的差距估计 (gapper estimator) 的方法来计算星系群的视线方向速度弥散 σ (与 Yang et al. (2007) 一致)。在此步我们只计算拥有三个及以上星系的星系群的 σ 值，由于此类星系只占星系群样本将近 20% 左右，故在“Morph- M_h ”样本中，同时拥有暗晕质量和 σ 值的星系从 35807 缩减为 8142 个，我们将此样本命名为“Morph- M_h/σ ”样本。在此步我们得到星系所处暗晕的质量 (M_h)，星系相对于星系群的退行速度 $|\Delta V|$ ，星系群视线方向的速度弥散 (σ) 以及星系所处暗晕的位置 (r/r_{180}) 等参数。在 3.4 节中我们将利用这些参数构建相空间图 (Phase Space Diagram, PSD)，对星系所处环境的动力学状态进行进一步分析。

值得一提的是，GSWLC-X2 数据集中计算恒星质量等参数时所用的初始质量函数 (Initial Mass Function, IMF) 为 Chabrier IMF (Chabrier, 2003)，所用宇宙学参数为 $\Omega_m = 0.272$ ， $\Omega_{\Lambda} = 0.728$ ， $h = 0.704$ ；而 Yang et al. (2007) 的星系群样本所用宇宙学参数为： $\Omega_m = 0.238$ ， $\Omega_{\Lambda} = 0.762$ ， $h = 0.73$ 。由于本文样本的平均红移仅为 $z \sim 0.1$ ，我们忽略两个样本中宇宙学参数 Ω_m 和 Ω_{Λ} 差异的影响。对于 h ，本文统一使用 $h = 0.73$ 对 GSWLC-X2 数据集中每个星系的 SFR 和恒星质量进行了改正。

星系的形态结构等信息来自 NSA 数据集。我们将“Morph- M_h -SFR”与 NSA 样本进行交叉匹配，获得 Sersic 指数、轴比、对称性等星系形态信息。匹配后的样本的数量仍为 31754 个。同时，我们将从 SDSS 官网得到的五个波段 (u、g、r、i、z) 的 model 星等与样本进行匹配。并且根据样本中星系的红移和位置信息所计算得到 Dnn (Distance with the Nearest Neighborhood) 参数，即每个星系距离最近星系的距离，单位为 kpc，用来表征星系收到的小尺度的环境效应。部分没有伴星系的源被舍去。将此样本命名为“Morph- M_h -SFR-Oth”样本，样本大小为 26503 个。

我们列出本工作所用样本及名称，考虑不同问题时将用不同样本进行分析，详见表 2-2。其中，只包含了棒相关信息的样本为“Morph”，样本大小为 56912 个。包含了棒相关信息和恒星形成相关信息的样本为“Morph-SFR”，样本大小为 43221 个。包含了棒相关信息和暗晕质量的样本为“Morph- M_h ”，样本大小为 35807 个。包含了棒相关信息，暗晕质量和暗晕速度弥散度星系的样本为“Morph- M_h/σ ”，样本大小为 8142 个。“Morph-SFR”和“Morph- M_h ”样本交叉所得样本为“Morph- M_h -SFR”，样本大小为 31754 个。“Morph-SFR”和“Morph- M_h/σ ”样本交叉所得样本为“Morph- M_h/σ -SFR”，样本大小为 7127 个。包含了棒相关信息、暗晕质量、恒星形成信息和其它综合物理参数信息的样本为“Morph- M_h -SFR-Oth”，样本大小为 26503 个。该样本将在第 4 章中用于棒特征和其它综合物理参数的相关性分析。

不同样本中无棒、弱棒和强棒三个子样本的数目也同样在表 2-2 中给出。基于以上样本，我们在进行对比研究的时候还需要进一步构建控制样本。在构建控制样本时，我们在对数空间对参数（比如恒星质量、暗晕质量）进行样本匹配，所选取阈值为 0.1 dex。

表 2-2 本文所用样本
Table 2-2 Sample

Name	strong bar	weak bar	no bar	Total
Morph	6723	16703	33486	56912
Morph-SFR	5151	12833	25237	43221
Morph- M_h	4848	10174	20785	35807
Morph- M_h -SFR	4273	9036	18445	31754
Morph- M_h/σ	1289	2513	4340	8142
Morph- M_h/σ -SFR	1125	2215	3787	7127
Morph- M_h -SFR-Oth	3587	7508	15408	26503

第3章 棒星系与恒星形成特征的统计分析

在本章中，我们分析棒星系与恒星形成特征的统计特征。在 3.1 节中我们介绍本文所使用的统计方法，在 3.2 节中我们研究棒星系沿恒星质量分布的比例，在 3.3 节中我们研究棒星系的 sSFR 分布情况，在 3.4 节中我们研究棒星系所处环境信息，在 3.5 节中我们控制环境来研究棒星系的 sSFR 分布情况。

3.1 统计中检验样本相关性算法

为了分析棒特征与 SFR、环境之间的相关性，我们需要用到统计学中检验样本相关性的算法。常见的算法有 K-S 检验 (Kolmogorov-Smirnov Test)，卡方检验 (Chi-Square Test)，T 检验 (T-Test)，F 检验 (F-Test) 等等。由于 T 检验只适用于数据服从正态分布且样本量较小 (一般 $n < 30$) 的情形，并只通过样本的均值来判断差异性；F 检验只通过样本的方差来判断是否有显著差异；卡方检验和 K-S 检验同作为应用广泛的非参数检验方法，区别在于卡方检验必须先将数据分组才能获得实际的观测频数，而对原始数据直接进行排序后就可以应用 K-S 检验，相对来说对数据的利用较为完整。故在本工作中我们用到 K-S 检验来作为检验样本相关性的工具，在本章 3.1.1 节中详细描述。在后续对星系的具体物理特征与棒的相关性进行排序时，我们用到了机器学习算法中的随机森林算法。在本章 3.1.2 节中进行详细描述。

3.1.1 K-S 检验

K-S 检验是一种重要的非参数检验方法，基于样本的累积分布函数 (Cumulative Distribution Function)，可以用来检验一个样本的分布是否遵从某种已知分布 (单样本 K-S 检验) 或检验两个样本的分布之间是否有显著差异 (两样本 K-S 检验)。在分析时，需要预先提出某种统计假设，称为零假设或原假设 (null hypothesis)。当原假设被肯定或否定后，即得到原结论或相反的结论。一般来说，原假设的内容为希望证明其错误的假设。例如在单样本 K-S 检验中，原假设为样本的分布与指定的假设分布无显著差异。这里指定的假设分布可以是正态分布、均匀分布、泊松分布等常用的统计分布。

该样本中变量 x 的累积分布函数为 $F(x) = P(X \leq x)$ ，表示随机变量 X 小于或等于某个值 x 的概率。然而累积分布函数描述的是理论上的概率分布，通常是未知的。实际应用中只能获得有限的样本数据，会导致我们无法知道总体的概率分布。因此我们只能通过基于实际观测数据的经验分布函数 (Empirical Distribution Function)，来估计累积分布函数。当样本量足够大时，根据大数定律 (当试验次数足够多时，随机事件的频率会接近概率)，经验分布函数会收敛到真实的累积分布函数，这意味着通过对样本数据的观测，我们可以获得对总体概率分布的较好估计。

首先将已知样本数据进行排序, 通过经验分布函数计算得到统计量 D 值, 表示两个分布函数之间的差值。经验分布函数计算公式以及 D 值计算公式如下:

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^n I(x_i \leq x), \quad I(x_i \leq x) = \begin{cases} 1 & x_i \leq x \\ 0 & x_i > x \end{cases} \quad (3-1)$$

$$D_n = \sup_x |F_n(x) - F_0(x)| \quad (3-2)$$

其中, $\hat{F}(x)$ 表示经验分布函数, $I(x_i \leq x)$ 为指示函数, 当 $x_i \leq x$ 时取值为 1, 否则取值为 0。 $F_n(x)$ 为被检验样本的经验分布函数, $F_0(x)$ 为假设分布的累计分布函数, \sup 表示 $|F_n(x) - F_0(x)|$ 的上确界, 即取所有 x 值上的两个分布函数之间的最大绝对差值。

然后根据 D 值和样本量计算得到 p 值。 p 值表示在原假设成立的情况下, 观察到 D 值及更极端情况下出现的概率。 Kolmogorov 分布表是一种用来查找 K-S 检验中统计量 D 值对应的 p 值的表格。它列出了不同的 D 值和相应的 p 值, 可以用来快速查找 p 值。但此方法只能提供近似的 p 值, 且精度有限。比较准确的 p 值可通过 R 或 Python 等语言中的包来计算。

原假设的成立与否, 即取决于 p 值和显著性水平。显著性水平一般取 0.05。当计算得到的 p 值小于此显著性水平 (或称为置信度) 时, 则拒绝原假设, 即认为该样本的分布不服从指定理论分布; 若 p 值大于此显著性水平, 则不能拒绝原假设, 即认为该样本的分布服从指定理论分布。

同理在两样本 K-S 检验中, 若样本分别为 X_1, X_2, \dots, X_m 和 Y_1, Y_2, \dots, Y_n 。两个样本的经验分布函数分别为 $F_{X,n}(x)$ 和 $F_{Y,m}(x)$, 则统计量 D 的计算公式为:

$$D = \sup_x |F_{X,n}(x) - F_{Y,m}(x)| \quad (3-3)$$

由于在我们的工作中, 我们想要得到不同棒星系样本之间的恒星形成性质的相关性, 所以我们应用两样本 K-S 检验并通过得到的 p 值来进行分析。后续在第 3.3 节中详细描述。

3.1.2 随机森林算法

在当前的科研工作中, 机器学习扮演着越来越重要的角色。它能帮助人们从大量的数据中提取有用的信息, 发掘潜在的规律和相关性。通过对已有数据进行分析和学习, 建立模型来找出数据间的关系, 并对新数据进行预测和分类。机器学习算法通常分为有监督机器学习算法 (Supervised Machine Learning Algorithm)、无监督机器学习算法 (Unsupervised Machine Learning Algorithm) 和半监督机器学习算法 (Semi-Supervised Machine Learning Algorithm)。有监督的一类是指在训练过程中, 所输入的数据集带有已知标签, 算法可以根据这些标签进行学习并建立

模型，建立好模型之后用来预测新数据的标签。有监督机器学习算法主要聚焦分类和回归两种问题，这两个问题对应的数据标签分别为离散值和连续值。常见的有监督的算法包括神经网络 (Ayyar et al., 2022)，随机森林 (Chi et al., 2023)，支持向量机 (Huertas-Company et al., 2008) 等等。无监督的一类则是指在输入训练数据集时并无已知的标签，需要算法自行发掘数据中的信息和规律。无监督机器学习算法主要聚焦降维和聚类两种问题。降维旨在通过减少数据的维度来提取关键信息量，降低分析的难度和成本 (de Souza et al., 2022)。聚类则是用模型去自行提取发现数据中的相似度，并将数据聚合成带有规律性的不同类别 (Fang et al., 2023)。半监督学习则介于有监督学习和无监督学习之间，利用大量的未标记数据和少量的有标签的数据，构建模型并预测新数据。在实际应用中由于相比于有监督学习，无需大量标签样本，同时相较无监督学习，又具有部分标签样本的优势，近期也得到了科研人员的关注 (Mong et al., 2023)。

随着天文领域中观测水平的发展，人们能看到更深更远的宇宙，获取到的天文数据也呈几何式增长。机器学习相关的算法也越来越多的应用在了天文领域当中，当前，机器学习算法在光变曲线分析 (Burhanudin et al., 2021)、测光红移 (Sadeh et al., 2016) 以及星系的形态学分析 (Fang et al., 2023) 上起到了不错的效果。

决策树算法属于有监督学习算法，可以用来做分类和回归问题。并且易于理解和实现，具有较好的可解释性。其利用树状结构进行判据，树状结构中有不同节点 (Node)，每个节点代表样本的不同特征。最初的节点为根节点 (Root Node)，中间部分的节点为决策节点 (Decision Node) 或子节点，而最末端的节点为叶子节点 (Leaf Node)，代表最后的决策值。节点与节点之间以分支 (Branch) 的形式进行连接。当数据输入时，决策树会根据每个节点所代表的特征来对数据进行判断，根据数据相应特征的值来决定下一个子节点的传递方向，以此类推，直至最后的叶子节点。具体可见图 3-1。

由于不同特征的选择以及在决策树上节点的排列顺序不同，会导致整体树状结构的不同，进而影响决策的精准性。所以如何选取特征以及设置特征的顺序也很关键。在每一个节点上的特征划分数据集的过程中，划分前后所带给数据的不确定性的变化，某种程度上意味着该特征的信息量的多少：划分前后数据不确定性的变化大，也就意味着该特征的信息量较多；划分前后数据不确定性的变化很小，也就说明该特征信息量较少。在决策树算法里，有两种参数可以计算特征的信息量，分别是熵 (Entropy) 和基尼系数 (Gini index)。用于分类问题的 ID3 (Iterative Dichotomiser 3) 决策树模型就用到了熵的概念来选择最优特征。熵是用来衡量一个节点或一个系统内不确定性的参数，其计算公式如下：

$$Entropy = - \sum_{k=1}^{|Y|} p_k \log_2 p_k \quad (3-4)$$

其中 Y 为该节点下类别集合， $|Y|$ 表示该类别集合中的类别个数， p_k 表示第

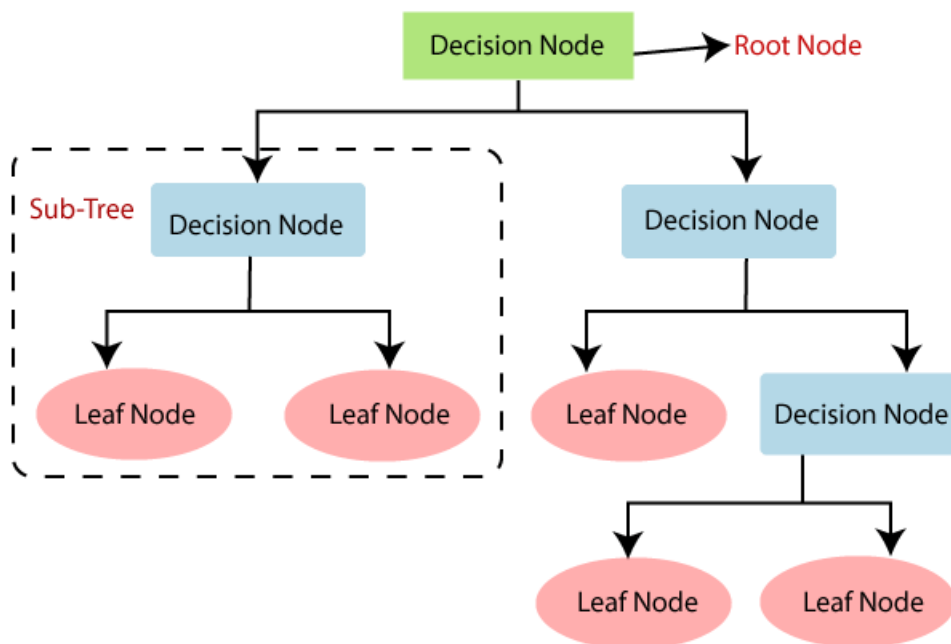


图 3-1 决策树流程图
Figure 3-1 Decision Tree

k 个类别里样本所占该节点样本总数的比例。熵越大代表节点或系统的不确定性越高，秩序越乱，反之则代表秩序越一致。以 2 分类问题为例，若两类中的样本数各占 50% 时，熵的值为 1；当只有其中一类时，熵的值为 0。也就是说，如果一个节点的熵越高，代表这个节点下的样本的一致性越低；如果一个节点的熵越低，代表这个节点下样本的一致性越高，更多的趋向于同一个分类。决策树算法的出发点即为尽快区分不同类别的样本，所以每个节点下的熵的值越小越好。故挑选特征时优先选取熵低的特征，并且将熵低的特征优先排列在较上方的节点，这样会更有效的完成决策。

而基尼系数同样为衡量节点或系统混乱度的参数，代表着数据的纯度。在既可以用于分类又可以用于回归问题的 CART (Classification and Regression Tree) 决策树模型中，就用到了基尼系数这一参数作为选择最优特征的指标。具体计算公式如下：

$$Gini = 1 - \sum_{k=1}^{|Y|} p_k^2 \quad (3-5)$$

与熵的计算公式一样，其中，Y 为该节点下类别集合，|Y| 表示该类别集合中的类别个数， p_k 表示第 k 个类别里样本所占该节点样本总数的比例。基尼系数越小，选择标准的纯度也就越高，对应更高的重要性。当 $Gini = 0$ 时，意味着当前节点内的分布绝对平均，所有值都分为一类；当 $Gini = 1$ 时，意味着当前节点内的分布极度不平均，所有值的分类都不一样。故一般会选择基尼系数最小的特征构建决策树。

在挑选完合适的特征后，我们需要确定决策树中节点的数量。节点数量会影响决策树的复杂度和泛化能力。当节点过少时，模型可能不够精确，导致预测不准确；当节点太多时，决策树可能会出现过拟合的情况，即决策树过于复杂，泛化能力下降。为了避免这种情况，可以采取剪枝的操作。剪枝就是指将决策树的某些子节点下面的节点删除，使决策树不过于复杂。除了剪枝之外，还有一种常用的方法来避免过拟合，那就是随机森林算法。

随机森林算法作为可以同时胜任有监督学习和无监督学习的一种算法，是决策树算法的一种集成，由无关联的多颗决策树组成。具体示意图见图 3-2。该示意图以三棵树为例，具体模型下树的棵数为超参数，可直接设定，随机森林通过构建多棵互不相关的决策树，随机抽取输入每棵树的数据集和所使用的样本特征，并将每颗决策树的结果以“少数服从多数”的原则输出，进而避免过拟合的情况，同时还增强了模型的泛化能力。以星系为例，星系的不同特征 (M 个) 被随机抽样 m 个放在决策树的节点 (m 个, $m < M$) 中，每棵树的 m 个特征都不尽相同；数据集中的星系被随机有放回抽出，Tree 1、Tree 2 和 Tree 3 所输入的数据集都不尽相同。将各自的数据集依次放进树最开始的节点，在节点进行判据时，决策树会根据星系该节点特征的数值 (连续值或离散值) 来将此星系分类并传递到不同的下个节点上，以此传递直至最后一个节点，并输出结果值。对于分类问题，如图 3-2 所示每棵树产生的结果按照多数原则，得到随机森林的最终结果；对于回归问题，每棵树得到的结果相加求平均作为最终结果。

随机森林算法作为可以用作解决分类，回归和聚类问题的算法，其具有准确，高效，并能处理高维数据的特点，并且可以纠正决策树过度拟合训练集的缺点。故在第 4 章中我们应用机器学习里的随机森林算法来对星系中的物理参数与棒的相关性进行分析。

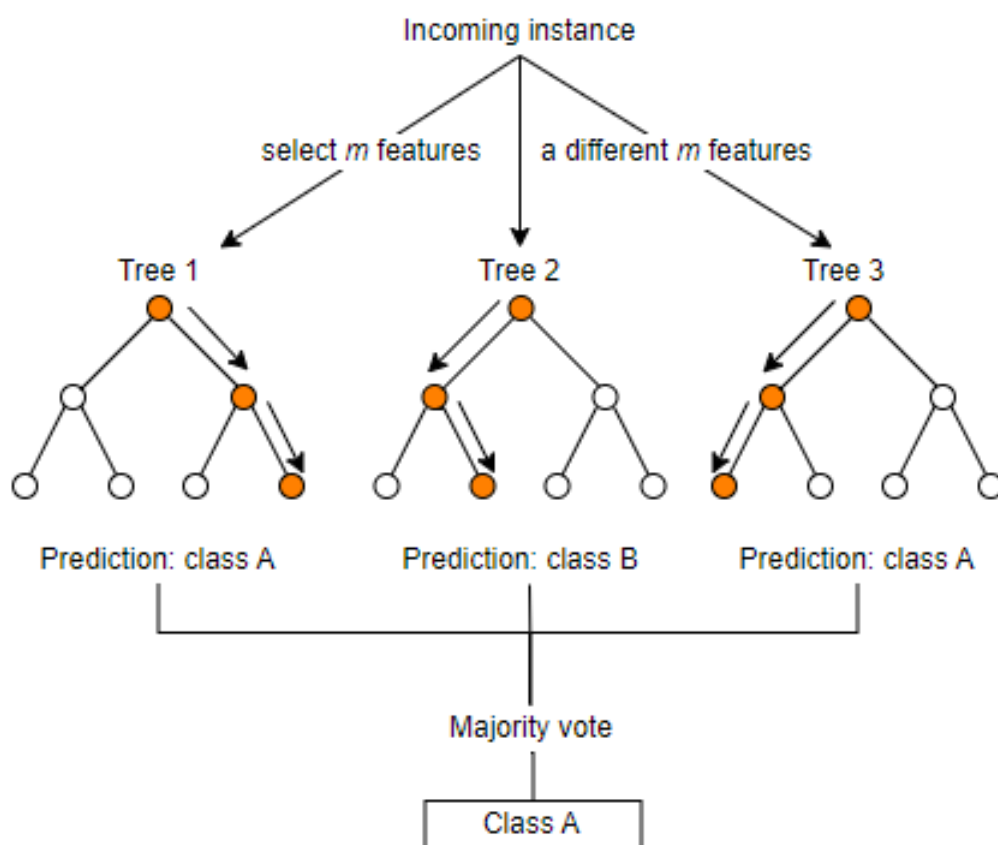


图 3-2 随机森林结构示意图
Figure 3-2 Random Forest

3.2 棒星系分布比例

上节中我们对所用的统计方法进行了概述，从本节开始我们进行具体的统计分析。首先，星系的恒星质量是星系的第一参量，与星系的其他参数存在明显的相关。故我们首先给出“Morph-SFR”样本中棒旋星系在不同恒星质量上的比例分布，见图 3-3。在对“Morph-SFR”样本沿恒星质量等数目分为十组后，我们计算在每组内三种棒星系样本所占比例。可以看出，强棒星系随星系恒星质量的增大，所占比例逐步增加，弱棒星系随星系恒星质量的增大，所占比例先下降后增加，在 $\log(M_*/M_\odot) = 10.4$ 位置处比例最低。而无棒星系样本则与弱棒星系样本趋势相反，在 $\log(M_*/M_\odot) = 10.4$ 位置处比例最高，随恒星质量增高比例降低。与前人工作一致 (Cervantes Sodi et al., 2015)。

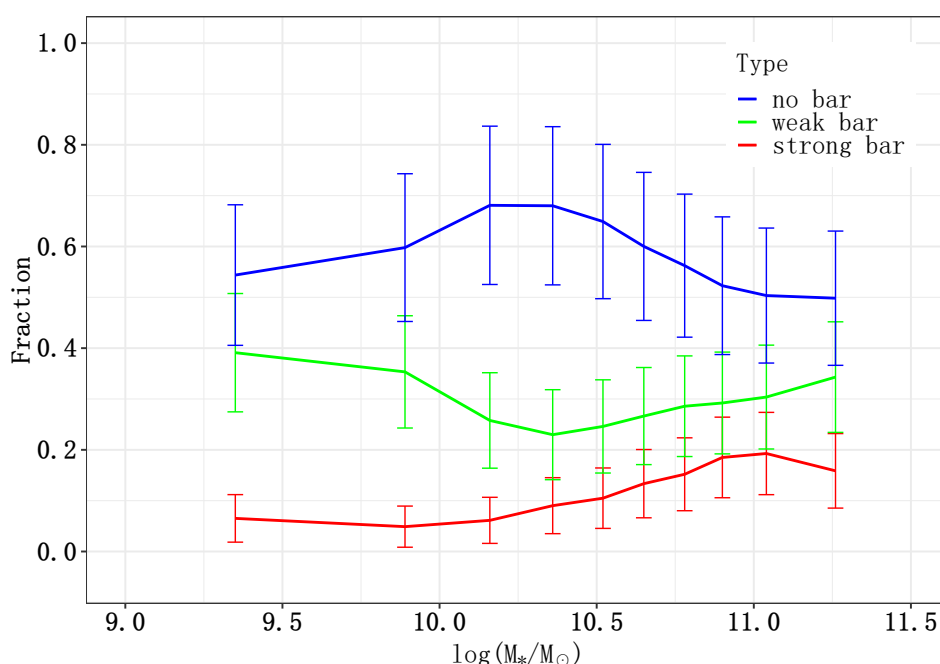


图 3-3 三种棒星系在不同恒星质量分布上的比例

Figure 3-3 The proportion of the three barred galaxies in the stellar mass distribution

3.3 棒星系与恒星形成

棒星系与恒星形成之间的相关性是我们研究的重点。首先，图 3-4 中给出“Morph-SFR”样本中的三种棒的恒星质量和比恒星形成率的分布（散点和等数密度轮廓），其横坐标为 $\log(M_*/M_\odot)$ ，纵坐标为 sSFR。红色数据为强棒样本 (SB)，绿色为弱棒样本 (WB)，蓝色为无棒样本 (NB)。上边界图为恒星质量数密度分布图，右边界图为 sSFR 数密度分布图。我们参照 Salim (2014) 的工作，将 $sSFR > -10.8$ 的星系定义为恒星形成星系 (“Star Forming” galaxy, SF)，将 $-11.8 < sSFR < -10.8$ 的星系定义为绿谷星系 (“Green Valley” galaxy, GV)，将 $sSFR < -11.8$ 的星系定义为熄灭星系 (“Quench” galaxy, Quench)。由上边际和右边际图中红色密度分布曲线可以看出，样本中强棒星系相对于弱棒星系和无棒星系，质量偏大，

sSFR 偏低。由于恒星质量和 sSFR 之间存在内禀的相关性，因此需要控制样本的恒星质量后，来进一步对比棒星系的 sSFR 特征。此外，由于本文重点考察的是星系的棒特征和 sSFR 的关系，而星系的颜色（特别是 $NUV - r$ ）直接体现的就是其 sSFR，因此我们在本节和后续分析都不对星系的颜色特征进行控制。

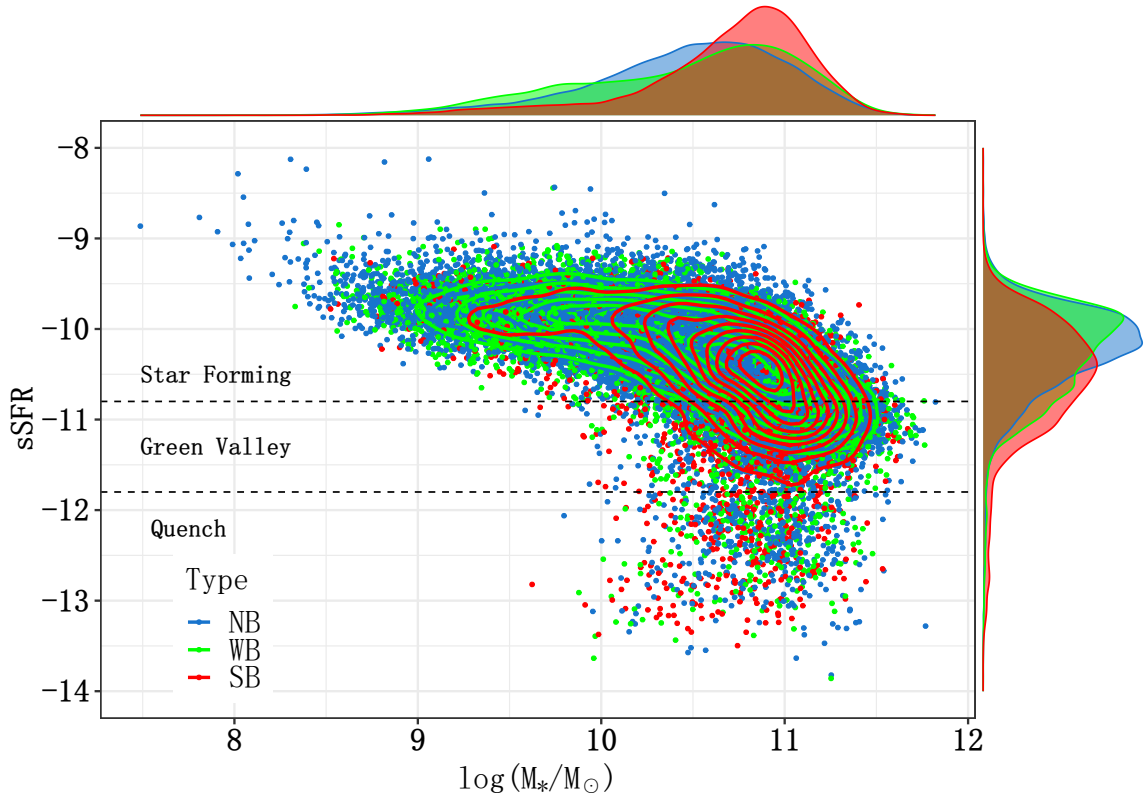


图 3-4 样本 “Morph-SFR” 中三种棒的恒星质量和比恒星形成率 (sSFR) 的散点图和轮廓图，其横坐标为恒星质量 ($\log(M_*/M_\odot)$)，纵坐标为比恒星形成率 (sSFR)。上边际图为恒星质量分布直方图，右边际图为 sSFR 分布直方图。红色图例为强棒样本 (SB:5151 个)，绿色为弱棒样本 (WB:12833 个)，蓝色为无棒样本 (NB:25237 个)

Figure 3-4 The scatter plot and contour plot of stellar mass and specific star formation rate (sSFR) of three types of bars in the sample "Morph-SFR". The x-axis is stellar mass ($\log M_*/M_\odot$), the Y-axis is sSFR. The upper marginal figure is histogram of distribution stellar mass, the right marginal figure is histogram of distribution of sSFR. The red color is strong bar sample (SB:5151), green color is weak bar sample (WB:12833) and the blue color is no bar sample (NB:25237)

基于强棒星系样本，我们通过阈值 (0.1 dex) 在无棒和弱棒样本中选取一一对应的恒星质量分布的星系来构建控制样本，最终得到三种类型星系的样本个数各为 5151 个。图 3-5 (a) 为控制恒星质量之后，三种棒样本的 sSFR 分布直方图。红色数据代表强棒，绿色代表弱棒，蓝色代表无棒。图中虚线为 $sSFR = -10.8$ ，右侧为 SF 星系，左侧为 Quench 星系和 GV 星系。可以看出棒旋星系相较于无棒星系，sSFR 分布更偏向于 sSFR 较低的一端，强棒星系更加显著。在图 3-5 (a) 的左上角图标中为三种棒样本中 sSFR 的中位值及误差，误差估计采用对样本进行有放回的随机抽样 1000 次后，计算得到的 1000 个中位值的标准差。由中位值可以看到有棒星系的 sSFR 相比于无棒星系更低，强棒星系最为明显。

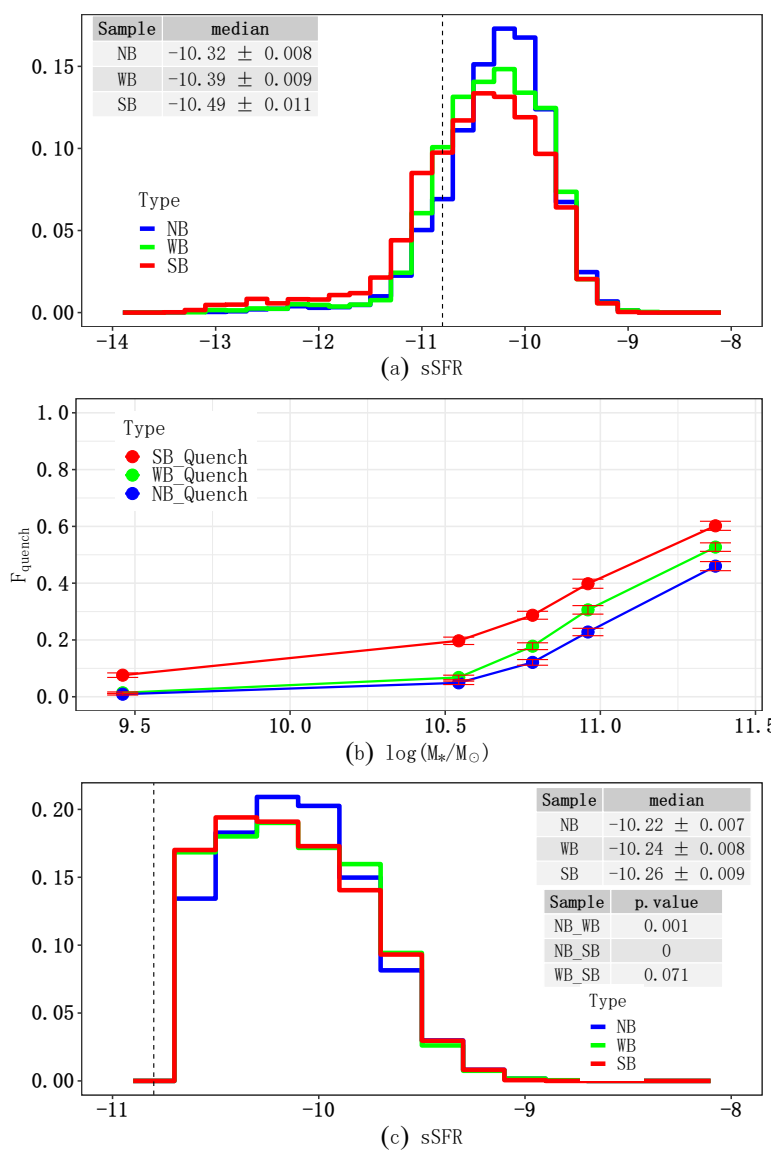


图 3-5 图 (a) 为控制了恒星质量 (M_*) 的三种棒的比恒星形成率 (sSFR) 的直方图。图 (b) 为控制了恒星质量 (M_*) 的三种棒样本中其 “Quench + GV” 星系占其样本的比例的点线图。图 (c) 为控制了恒星质量 (M_*) 的三种棒中恒星形成星系部分的比恒星形成率 (sSFR) 的直方图。红色图例代表强棒，绿色代表弱棒，蓝色代表无棒

Figure 3-5 Figure (a) is a histogram of the sSFR of the three types of bars which controlled stellar mass. Figure (b) is a point-line diagram of the proportion of “Quench + GV” galaxies in each bars samples with stellar mass controlled. Figure (c) is a histogram of the sSFR of Star-Forming part of the three types of bars which controlled stellar mass. The red color is strong bar sample, the green color is weak bar sample and the blue color is no bar sample

由于熄灭星系整体比例并不高，所以在研究 sSFR 整体分布中可能并不能显著显示。因此，我们引入熄灭星系所占据的比例 (Quench Fraction, F_{quench}) 这一参数来进一步分析。其计算公式如下：

$$F_{quench} = \frac{N_{quench} + N_{GV}}{N_{Total}}. \quad (3-6)$$

其中，N 为不同类型的星系个数。图 3-5 (b) 给出了控制样本中三种棒星系中各自的 F_{quench} 。我们将各样本等数目分为 5 个间隔，每个间隔中所有星系的恒星质量的均值为此间隔在 x 轴的横坐标。 F_{quench} 的误差同样是通过有放回的随机抽样进行估算。可以看到，随着星系恒星质量的增大，各样本中的 F_{quench} 都随之增加 (Donnari et al., 2021)。但是对于强棒、弱棒和无棒样本， F_{quench} 存在显著的系统差异，强棒星系的 F_{quench} 显著而且系统性的高于弱棒和无棒星系。弱棒和无棒星系中的 F_{quench} 整体上比较接近，但弱棒星系仍然系统性的高于无棒星系。

图 3-5 (c) 给出了控制样本中三种棒星系内 SF 星系部分的比恒星形成率 (sSFR) 分布。其中无棒星系 SF 部分的数目为 4260 个，弱棒星系 SF 部分的数目为 4020 个，强棒星系 SF 部分的数目为 3545 个。此外，我们将中位值和 KS 检验的 P 值列在图 3-5 (c) 的右上角图标中，可以看到三种棒星系样本中的恒星形成部分的 sSFR 的中位值基本一样，由 P 值可以看到弱棒和强棒样本中，恒星形成 (SF) 星系的 sSFR 较近似服从同一分布。

3.4 棒星系与环境

星系所处的环境因素，是否会对我们上节的结果产生影响尚不清楚。在本节中，我们对棒星系所处环境参量进行分析，其中具体参数包括星系所处星系群的暗晕质量 (M_h)，归一化后的星系所处星系群的位置 (r / r_{180})，以及由星系群速度弥散归一化后的星系的视线方向退行速度 ($|\Delta V| / \sigma$)。我们首先画出“Morph- M_h -SFR”样本中控制恒星质量的三种棒样本的暗晕质量直方图，见图 3-6(a)。其中红色数据代表强棒 (SB)，绿色代表弱棒 (WB)，蓝色代表无棒 (NB)，可以明显看到棒强度越强的样本，其星系所处星系群的暗晕质量分布就越趋向于大暗晕端。此外，通过右上角图标中暗晕质量的中位值，也可以看到有棒星系更偏向于分布在大暗晕里，强棒星系最为显著。

暗晕质量不是描述星系环境的唯一参数，星系在暗晕中所处位置及速度状态也很重要。相空间图 (Phase-Space Diagrams, PSD) 在研究星系团的形成历史 (Muzzin et al., 2014) 以及回溯星系在其中的轨道历史等方面是一种有效的工具。相空间图的横坐标为归一化后的星系所处暗晕的位置 (r / r_{180})，纵坐标为星系团速度弥散归一化后的星系的视线方向退行速度 ($|\Delta V| / \sigma$)，通过星系在相空间图上分布位置的不同 (Jaffé et al., 2015)，可以将星系团中的星系所处位置分为位力化区域 (Virialized) 和内落区域 (Infalling)。位力化区域中的星系一般

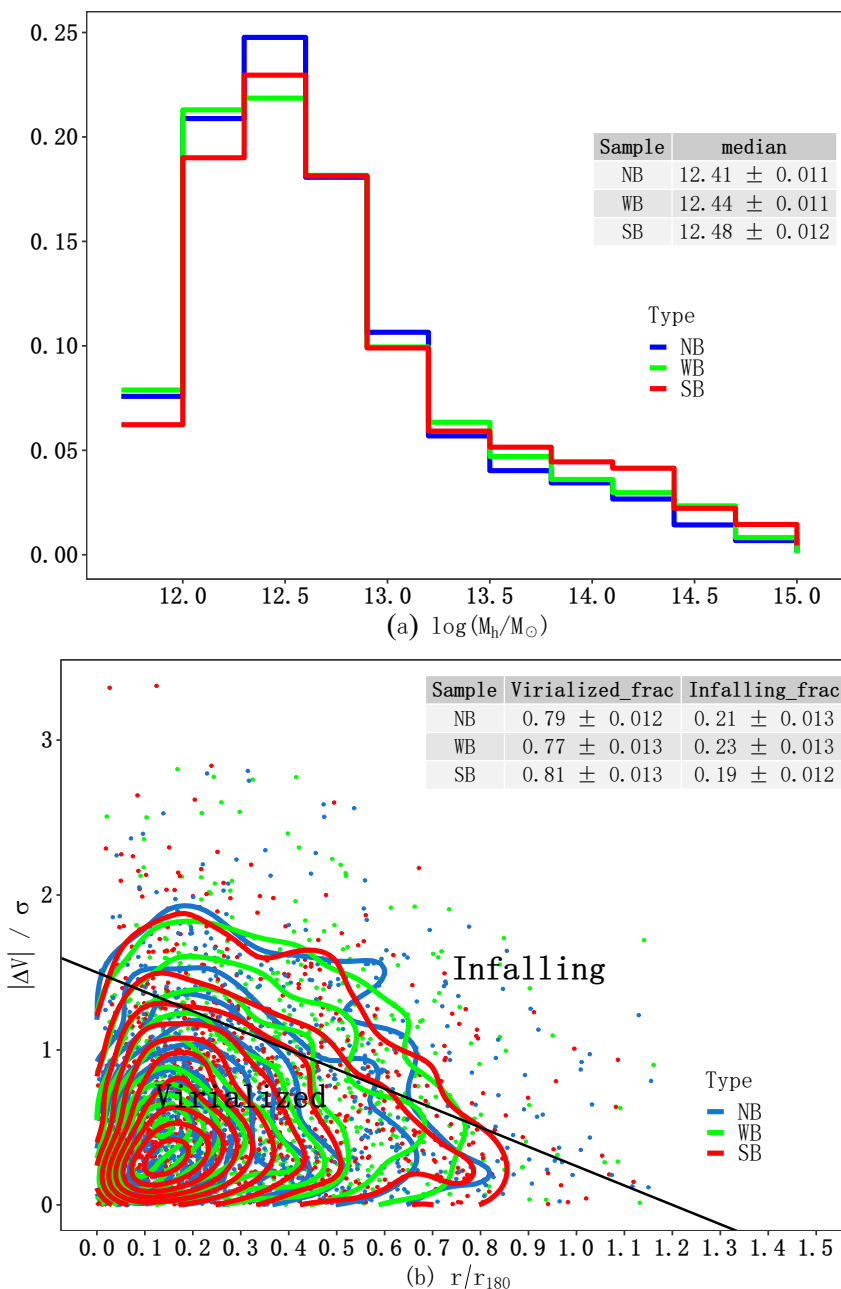


图 3-6 图 (a) 为 “Morph- M_h -SFR” 样本中，控制了 M_* 的三种棒的 M_h/M_\odot 直方图。三种棒的样本个数各为 4273 个。图 (b) 为 “Morph- M_h/σ -SFR” 样本中，控制了 M_* 和 M_h/M_\odot 的三种棒星系的相空间图。图中黑线为 Jaffé et al. (2015) 的方法。三种棒的样本个数各为 1031 个。红色图例代表强棒，绿色代表弱棒，蓝色代表无棒

Figure 3-6 Figure (a) is the dark halo mass histogram of the three types of bars which stellar mass controlled in sample “Morph- M_h -SFR”. The number of each bar sample is 4273. Figure (b) is the phase-space diagram of the three types of bars sample which stellar mass and halo mass controlled in sample “Morph- M_h/σ -SFR”. The boundary of “Virialized” and “Infalling” region comes from Jaffé et al. (2015). The number of each bar types is 1031. The red color is strong bar sample, the green color is weak bar sample and the blue color is no bar sample

认为星系已经在星系团内部并且在动力学状态下处于动态平衡的状态。而内落区域表示星系正在处于从外围掉落到星系团内部的一个过程。那么，相空间图所表征的不同状态是否会和暗晕质量一样，与棒的强度具有相关性？我们选取“Morph- M_h/σ -SFR”星系样本，控制恒星质量和暗晕质量之后（阈值都设为 0.1 dex），得到三种棒的样本个数各 1031 个。在样本控制过程中，部分强棒星系在阈值范围内未获得样本匹配，因此最终样本数目有所下降。我们在 3-6(b) 给出相空间图，其中红色数据代表强棒 (SB)，绿色代表弱棒 (WB)，蓝色代表无棒 (NB)。在划分相空间区域时，我们采用 Jaffé et al. (2015) 的方法，如图中黑线所示，所用公式如下：

$$|\Delta V|/\sigma \leq 1.5 - 1.25r/r_{180}, \quad r/r_{180} \leq 1.2. \quad (3-7)$$

在相空间图上，三种类型星系的分布存在细微差别，其中强棒，弱棒和无棒中处于位力化区域的比例分别为 0.81, 0.77 和 0.79。从比例数据上来看，处于位力化区域的强棒星系比例是最高的，这与强棒星系整体上处于偏向高密环境（处于大质量暗晕内）的趋势是一致的。但是，有趣的是，处于位力化区域的弱棒星系的比例是最低的，而不是无棒星系。此外，考虑统计误差，三种星系的比例之间的差异也并不显著。因此，要更深程度确认三种类型星系在相空间分布图上可能存在的差异，还需要进一步扩大棒星系的样本。

在本节中，通过棒星系样本的环境特征分析，我们得到，有棒星系和无棒星系所处的环境之间（特别是在暗晕质量上）存在系统差别，其中强棒星系明显偏向于高密环境的结论。由于高密环境整体上对星系的恒星形成起抑制作用，因此我们在 3.3 节中得到的有棒星系整体上具有更高比例的熄灭星系及恒星形成星系 sSFR 系统性偏低的结论中需要进一步考虑环境效应可能带来的偏差。

3.5 控制环境因素后棒星系的恒星形成

由上节可知，对于分析棒星系的恒星形成信息，考虑环境因素是不可或缺的一步。本节中，我们基于“Morph- M_h/σ -SFR”样本，并同时控制恒星质量和暗晕质量。此外我们通过分离开位力化星系和内落星系，来消除星系位于相空间图的不同位置中可能带来的环境效应的偏差。最终，我们得到位于位力化区域的三种棒星系的控制样本各 786 个，位于内落区域的三种棒星系数目各 138 个。在样本控制过程中，由于样本匹配分别在位力化区域和内落区域进行，因此在相同的质量和暗晕质量阈值下，最终匹配出的样本 (786+138) 相比于 3.4 节中控制样本数 (1031) 有进一步下降。

我们在图 3-7 中的左侧给出位力化区域星系的性质，内落区域星系相应位于右侧图 (a) (b) 为 sSFR 分布直方图，左上角图标中列出三种棒星系样本中 sSFR 中位值及误差。由分布直方图及中位的 sSFR 值可以看到，处于不同相空间图的位置，强棒星系相较于弱棒和无棒星系，其平均的 sSFR 都更低，且位力化区域

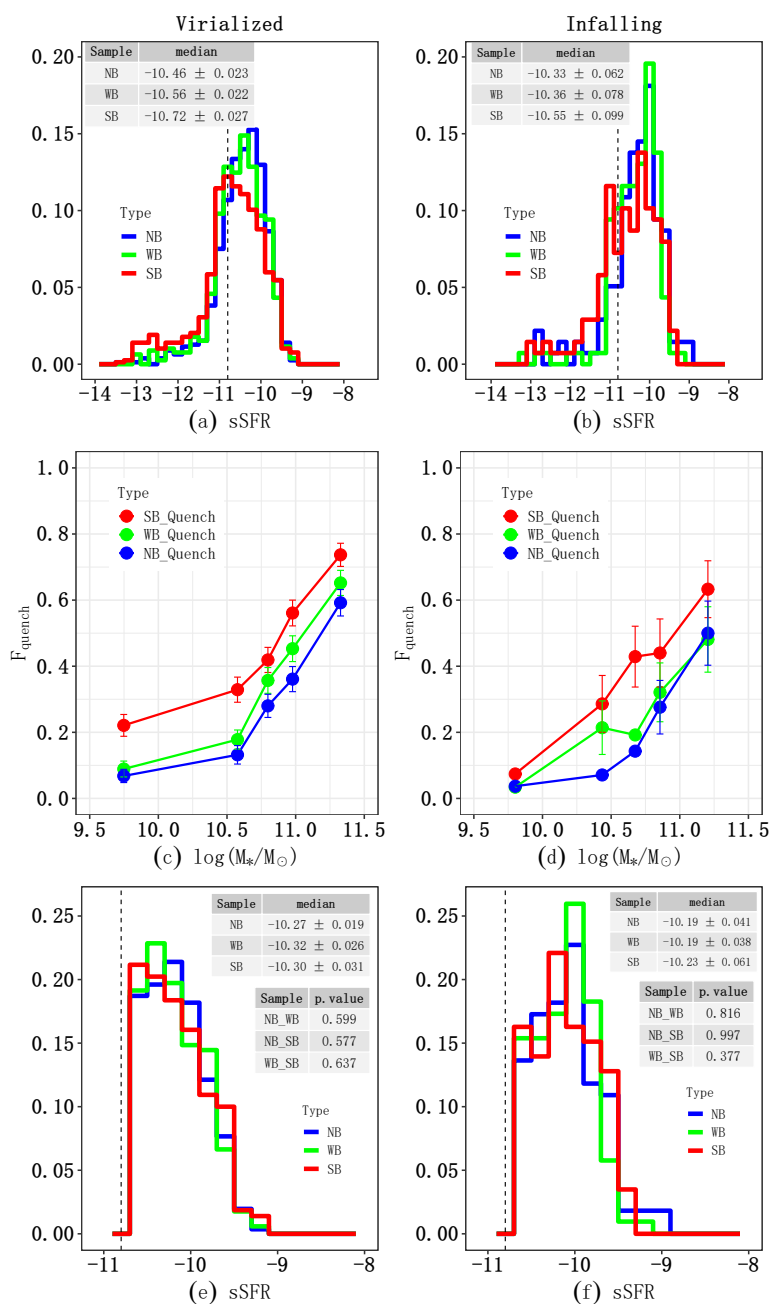


图 3-7 左列为位力化星系样本，右图为内落星系样本。(a)(b) 为 sSFR 的分布直方图 ((a): 各 786 个星系, (b): 各 138 个星系), (c)(d) 为不同棒样本中, 其”Quench”+ “Green Valley” 星系占其样本的比例的点线图 ((c): 各 786 个星系, (d): 各 138 个星系)。(e)(f) 为三种棒星系样本中 SF 部分 sSFR 的分布直方图 ((e): 共 1503 个, (f): 共 300 个)。图中控制了恒星质量和暗晕质量

Figure 3-7 The left column is Virialized region galaxies and the right column is Infalling region galaxies. The upper panel is the histogram of sSFR.((a): each bar sample has 786 galaxies. (b): each bar sample has 138 galaxies.) The middle panel is the proportion of “Quench + GV” galaxies in each bar samples.((c): each bar sample has 786 galaxies. (d): each bar sample has 138 galaxies.) The lower panel is the histogram of sSFR of SF part. ((e): totally 1503 galaxies. (f): totally 300 galaxies.)The samples of different panel of figure are all controlled by stellar mass and halo mass

这种差异更显著。(c)(d) 两图为不同棒样本中，其“Quench + GV”星系占其样本的比例。每个点的误差同图3-5(b)，是通过有放回随机抽样计算所得。可以看到，在不同的细分环境中，有棒星系中的 F_{quench} 都高于无棒星系，而且强棒高于弱棒。也就是说，在同时控制质量和环境因素之后，棒特征越为明显的星系中包含更多比例的“熄灭”星系。到此为止，我们给出了星系中的棒特征和恒星形成特征直接相关的证据。但是，这种更高比例的“熄灭”星系是由什么样的物理机制造成的呢？是因为星系中的棒特征在某种物理机制（如降低恒星形成效率）作用下降低了星系 sSFR，从而使得星系进入熄灭状态？或者已经恒星形成“熄灭”的星系中存在某种额外的机制更容易形成棒的特征？为了区分这两种场景，我们进一步考察恒星形成星系的 sSFR 和棒特征的关系 (e)(f) 两图给出了控制样本中三种棒星系内 SF 星系部分 (sSFR > -10.8) 的 sSFR 分布。图 (e) 为位力化区域星系的 SF 部分，三种棒星系共 1503 个。图 (f) 为内落区域星系的 SF 部分，三种棒星系共 300 个。此外，我们将这些星系 sSFR 的中位值标记在右上角中。由 sSFR 的中位值可以看到，在控制了恒星质量和环境因素之后，当只考虑 SF 星系部分 (sSFR > -10.8) 时，三种不同棒特征星系的 sSFR 不再具有显著差别。为了进一步说明这一点，我们对这个三个样本的 sSFR 的分布进行了两两比较的 KS 检验，并将检验结果中表示样本两两对比是否具有相同分布的 P 值标记在图的右上角。由 KS 检验的 P 值可以看到，在只考虑 SF 星系时，和我们预期棒特征能整体上降低星系的 sSFR 的猜测不同的是，三种不同棒特征的 SF 星系样本具有相同的 sSFR 分布。

综上，结合上中下三行图例可以看到，在图 (a)(b) 给出的有棒星系的平均 sSFR 系统偏低的原因是因为有棒星系中有更高比例的“熄灭”星系 (图 (c)(d))。而只考虑 SF 星系的话，在考虑质量和环境的选择偏差之后，有棒和无棒星系的恒星形成特征并没有显著差别。这正是本章得到的最主要的结论。

根据以上结论，可以推断，如果是棒特征导致了“熄灭”星系比例偏高的话 (图 (c)(d))，那么这个棒特征驱动的恒星形成熄灭过程必然是一个比较迅速的过程，只有这样才能在观测上不影响 SF 星系的 sSFR (图 (e)(f))。如果对于具有棒特征的“熄灭”星系来说，并不是因为棒导致了恒星形成的“熄灭”，而是反过来“熄灭”星系中存在某种更容易激发棒特征的假设，这样的因果关系也是可能的。值得指出的是，我们的统计研究本质上只是给出两种物理特征之间相关性，而并不能给出其因果性。要厘清两种不同的因果关系，需要借助数值模拟等其它研究工具 (Cavanagh et al., 2022; Rosas-Guevara et al., 2022)，而这已经超出本文的研究范围。

第 4 章 棒特征与其它物理参数的关系

在第 1 章中我们提到星系的内外两种物理参数存在一定的相关性, 在第 2 章中我们介绍了所用的数据集、各数据集所能提供的信息以及本文所用样本。在第 3 章中我们介绍了本文所用的统计方法, 分析了星系的棒特征与星系的内外两种物理参数的相关性。但值得注意的是, 第 3 章中只涉及到星系的恒星形成性质和星系所处的环境特征, 并没有涉及到其它参量, 例如星系的大小, 星系的形态参数, 星系的颜色等等。且由于这些参数互相之间存在复杂的相关性 (如可利用恒星质量-恒星形成率之间的相关性, 来区分熄灭星系、绿谷星系和恒星形成星系三种类别的星系), 不同参数之间或强或弱的相关关系使分析单一变量对棒特征的影响变得非常困难。以往所用的简单的线性相关方法很难分离出某类变量或某几类变量对棒特征的影响高低, 故我们采用数据驱动的机器学习算法来对星系中的棒特征进行分析。有借鉴意义的工作包括 [Teimoorinia et al. \(2016\)](#) 应用了人工神经网络 (Artificial Neural Network, ANN) 模型, 对多个星系物理性质在星系的熄灭过程中起到的重要性进行排序, 发现与星系中心质量相关的参数 (例如中心速度弥散度、核球质量、B/T 即核球质量与核球加盘的质量之比) 起到更重要的作用; [Curti et al. \(2022\)](#) 应用了 ANN 和随机森林两种机器学习方法, 来分析哪些物理性质更能帮助预测本地的恒星形成星系在 BPT 图上的位置。

综上, 我们不仅希望能找到一个可以组合各种星系特征来综合预言棒特征的方法, 更希望找到这些特征中的主导因素。例如具体的某种星系参数会较大的影响星系中的棒特征, 这种数据驱动寻找发现的主导因素将有助于我们更好的发现和理解与棒现象相关的物理过程, 且有效的帮助后续的棒相关研究。在本章节中, 我们选取 “Bar- M_h -SFR-Oth” 样本中的数据以及部分物理参数, 通过训练集来对 CART 随机森林分类模型进行训练, 并用模型对样本中的测试集进行预测, 同时根据模型中每个特征的 Gini 系数来分析不同特征与棒的相关性次序。

4.1 模型所用样本及物理参数

我们选取了 “Morph- M_h -SFR-Oth” 样本, 样本量为 26503 个, 其中强棒星系 3587 个, 弱棒星系 7508 个, 无棒星系 15408 个。同时, 在数据集中我们筛选了 12 个物理参数来分析, 具体如下:

- 来自 GSWLC-X2 数据集的 M_* 和来自 GZD 数据集的 R_{50} 、 R_{90} 等参数表示星系的质量和星系的物理尺度;
- 来自 NSA 数据集的 b/a (轴比)、Sersic 指数、Asymmetry (对称性) 等参数表示星系的具体形态结构;
- 来自 SDSS 官网的 g-r 颜色, 来自 GSWLC-X2 数据集的 SFR、sSFR 等参数表示星系中恒星形成的状态;

- 来自星系群样本的 M_h 、 r/r_{180} 和本工作算出的 Dnn (可见 2.5 节) 等参数表示星系的环境状态。

4.2 随机森林模型

在样本方面, 为将模型训练为二分类模型, 同时也为分析弱棒星系性质, 我们分别将弱棒星系样本划分给强棒星系样本和无棒星系样本, 用此两种划分方法得到的样本, 分别进行随机森林模型的训练和预测。为方便起见, 在弱棒星系和无棒星系划分在一起时, 将弱棒星系和无棒星系样本统一标签为“非强棒星系”, 强棒星系对应标签为“强棒星系”, 该样本所训练随机森林模型具体可见 4.2.1 节。在弱棒星系和强棒星系划分在一起时, 将弱棒星系和强棒星系样本统一标签为“有棒星系”, 强棒星系对应标签为“无棒星系”, 该样本所训练随机森林模型具体可见 4.2.2 节。

在特征方面, 我们用 4.1 节中的物理参数来训练随机森林模型, 该模型用来判断所输入的星系样本棒的信息: 其输入为星系样本参数信息, 输出为星系关于棒的判据。在随机森林模型中我们设置决策树的棵树为 100 颗, 每棵树的节点数目为 5 个, 以有放回的随机抽样的形式抽取数据放进模型中进行训练。同时, 我们应用随机森林模型输出星系中相关参数的对棒结构的 Gini 系数, 并对其进行排序, 由此分析哪些参数与棒的相关性更紧密。

4.2.1 “强棒星系”和“非强棒星系”

棒星系样本总数为 26503 个, 在将弱棒样本归为无棒样本后, “强棒样本”和“非强棒样本”的数目分别为 3587 和 22916, 低于 1: 6 的比例。在将样本以 7: 3 的比例分为训练集和测试集后, 训练集和测试集的样本数目分别为 18552 和 7951。在训练集中, “强棒样本”和“非强棒样本”的数目分别为 2537 和 16015。即在我们的样本中, 由于强棒样本、弱棒样本和无棒样本的数量相差过大, 导致出现了样本不平衡的问题。直接带入随机森林模型会导致模型在训练时过多学习的是多数的样本, 无法充分学习少数的样本。在预测时会导致偏向于多数的类, 预测少数类时性能较差。

有两种方法可以解决数据不平衡的问题, 分别是上采样/过采样 (over-sampling) 和下采样/欠采样 (under-sampling)。上采样是指在不平衡数据集中增加少数样本数量的方法, 可以通过随机复制现有的少数类别的样本或生成新的少数类别样本来实现。上采样可以帮助平衡数据集中的各个类别样本数量, 但由于复制样本的过程, 可能会导致过拟合的情况。下采样是指在平衡样本时, 减少多数类别的样本数量的过程, 这可以通过随机删除多数类别的样本或使用聚类算法来挑选代表性的多数类别的样本来实现。但同样由于删除或扔掉了部分样本, 可能会导致信息的丢失。在后续分析时我们将分别应用这两种方法, 并对两种结果进行比较。

在分析中我们用到了混淆矩阵, 该矩阵是一种常用的分类模型评估工具, 它

显示了模型在预测正类和负类时的表现，其中，横坐标代表真实值，纵坐标代表预测值。左上方为 True Positive(TP), 代表模型预测正确的正类样本数；右上方为 False Positive(FP), 代表模型预测错误的正类样本数；左下方为 False Negative(FN), 代表模型预测错误的负类样本数；右下方为 True Negative(TN), 代表模型预测正确的负类样本数。颜色的深浅代表样本数的多少，颜色越深代表样本数越多。

准确率 (Accuracy) 是分类模型中常用的评估指标之一，它表示模型正确预测的样本数占总样本数的比例，即模型预测正确的正类样本数和模型预测正确的负类样本数，占总预测样本数的比例。其计算公式为：

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4-1)$$

首先我们采用下采样的方法，将训练样本中，“非强棒样本”的数目降为和“强棒样本”数目一样的 2537 个，放入随机森林模型中进行训练。我们的测试集样本占棒星系样本的 30%，样本数为 7951，其中“强棒样本”和“非强棒样本”分别为 1071 个和 6880 个。我们用测试集样本放入随机森林模型进行测试。得到的混淆矩阵 (Confusion Matrix) 如图 4-1 所示。

故根据混淆矩阵计算出的准确率为 0.719。同时将每个特征值的 Gini 系数进行归一化并输出，视为每个特征与棒的相关性，如图 4-2 所示。

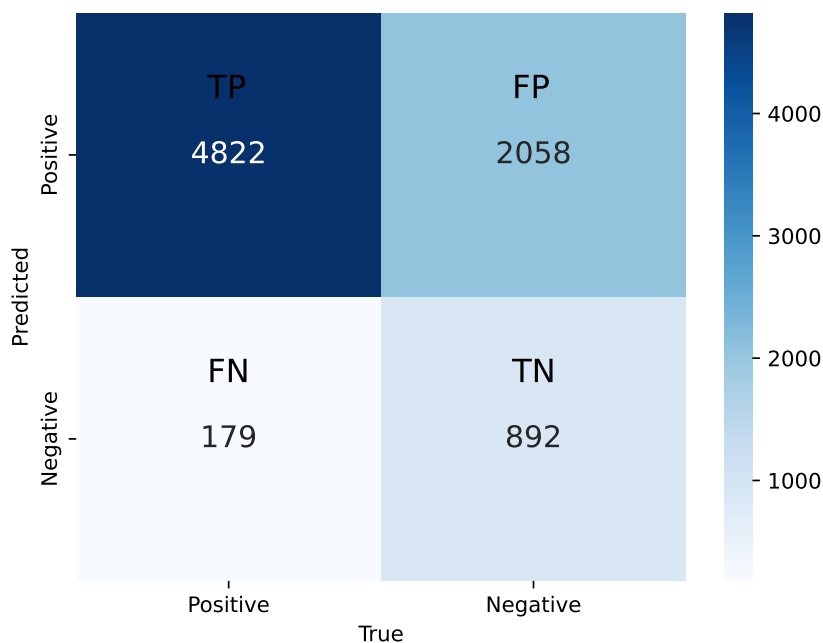


图 4-1 “强棒星系”和“非强棒星系”，下采样方法中的随机森林模型中的混淆矩阵
Figure 4-1 The confusion matrix of the random forest model with the with the down-sampling method of “strong bar” and “non-strong bar” galaxies

同时我们也应用上采样的方法来去除数据不平衡的影响，通过对“强棒样本”进行随机复制来增加其样本数量到和“非强棒样本”相同的 16015 个，放入

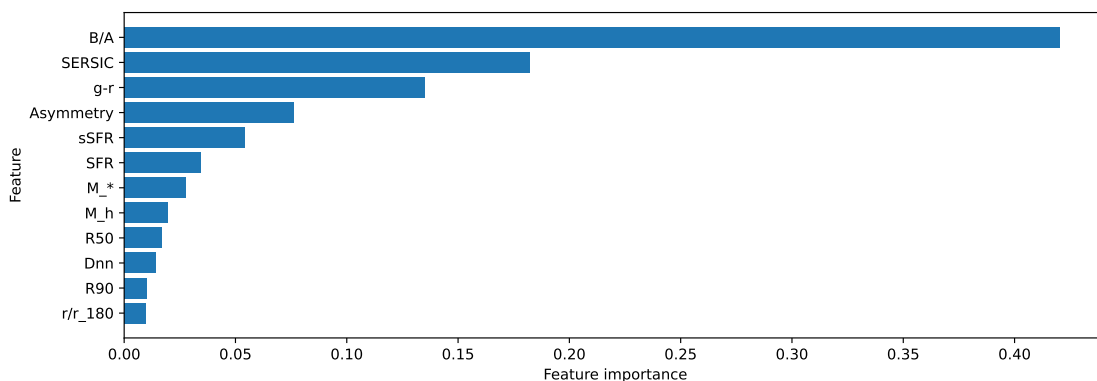


图 4-2 “强棒星系”和“非强棒星系”，下采样方法中的随机森林模型得到的参数重要性排序
 Figure 4-2 The ranking of feature importance obtained from the random forest model

随机森林模型中进行训练。与应用下采样方法时一样，我们的测试集样本占棒星系样本的 30%，样本数为 7951，其中“强棒样本”和“非强棒样本”分别为 1071 个和 6880 个。我们用测试集样本放入随机森林模型进行测试。得到的混淆矩阵如图 4-3 所示。根据混淆矩阵得到的准确率为：0.719。同时将每个特征值的 Gini 系数进行归一化并输出，视为每个特征与棒的相关性，如图 4-4 所示。

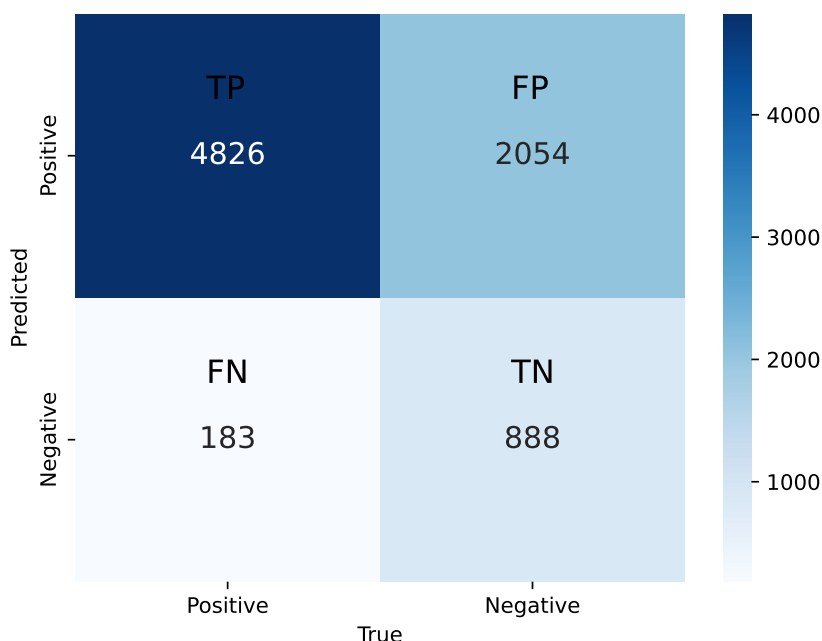


图 4-3 “强棒星系”和“非强棒星系”，上采样方法中的随机森林模型中的混淆矩阵
 Figure 4-3 The confusion matrix of the random forest model with the up-sampling method of “strong bar” and “non-strong bar” galaxies

在这里可以看到，上采样和下采样的结果相差不大，仅体现在极少部分的预测差别上。根据 Gini 系数可以看出，在划分“强棒星系”和“非强棒星系”的随机森林模型中，星系的轴比，Sersic 指数和 g-r 颜色重要性较高，重要性系数

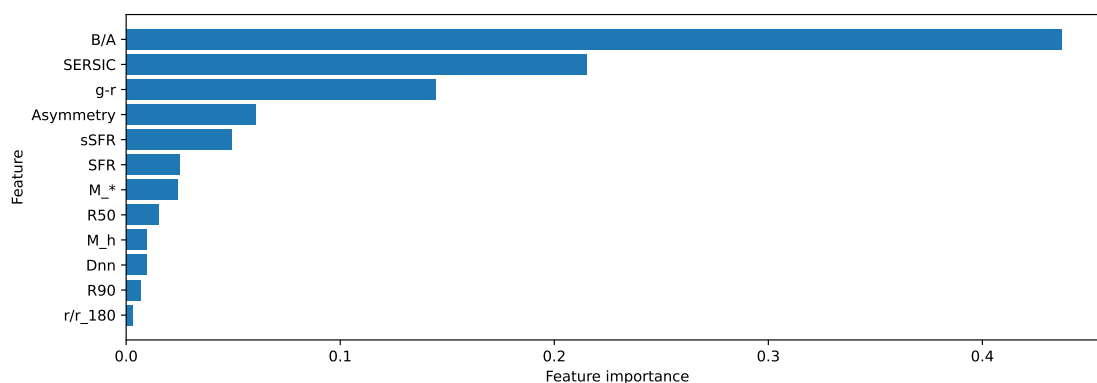


图 4-4 “强棒星系”和“非强棒星系”，上采样方法中的随机森林模型得到的参数重要性排序
Figure 4-4 The ranking of feature importance obtained from the random forest model

都高于 0.1，其余参数重要性都较低。“强棒星系”和“非强棒星系”中这三种参数的分布直方图可见图 4-5。可以看到“强棒星系”样本在 b/a 上的分布更偏向于偏小的一端，而“非强棒星系”偏向于 b/a 较大的一端。在 Sersic 指数的分布上，“强棒星系”样本更偏向于较大的 Sersic 指数，“非强棒星系”样本更偏向于较小的 Sersic 指数。在 $g-r$ 分布上，“强棒星系”样本偏向于更大的 $g-r$ （即更红）上，而“非强棒星系”则相反。

其中轴比的相关性最强，其原因可能是两方面的。一方面，带棒星系由于棒的存在，可能使得测量所得 b/a 偏向小的数值。另外一方面，也可能是星系的图像一定程度上影响志愿者的观察和投票，轴比较小的星系（细长型）可能容易使志愿者更多的选择投票为带棒的星系。值得一提的是在 GZD 决策树中判断为侧向于观察者 (edge-on) 的星系，并不会交予志愿者去做棒相关的判断 (参见图 2-4)。因此，是什么原因导致的 b/a 和棒特征相关，还需要我们在后续工作中继续开展相关的详细分析和研究。Sersic 指数作为拟合星系面亮度轮廓的参数，其与棒的相关性大的原因既可能是棒旋星系中棒使气体内流的效应的体现，也可能是 Sersic 指数偏大的星系中更容易触发棒特征。而 $g-r$ 颜色相关性大，与前人结论一致 (Kim et al., 2017; Consolandi et al., 2017)，原因可能是强棒星系有较多的“熄灭”星系的比例。

其余参数重要性的排列如下：Asymmetry、sSFR、SFR、 M_* 、 R_{50} 、 M_h 、Dnn、 R_{90} 和 r/r_{180} 。但是，整体贡献的相关性都小于 0.1。这些弱相关性可能涉及非常复杂的物理过程，本文就不再一一讨论。

4.2.2 “有棒星系”和“无棒星系”

棒星系样本总数为 26503 个，在将弱棒样本归为强棒样本后，“有棒星系”和“无棒星系”的数目分别为 11095 个和 15408 个，接近于 5:7 的比例，故在该随机森林模型中，无需进行上采样或下采样的操作。在将样本以 7:3 的比例分为训练集和测试集后，训练集和测试集的样本数目分别为 18552 个和 7951 个。在训练集中，“有棒星系”和“无棒星系”的数目分别为 7794 个和 10758 个。

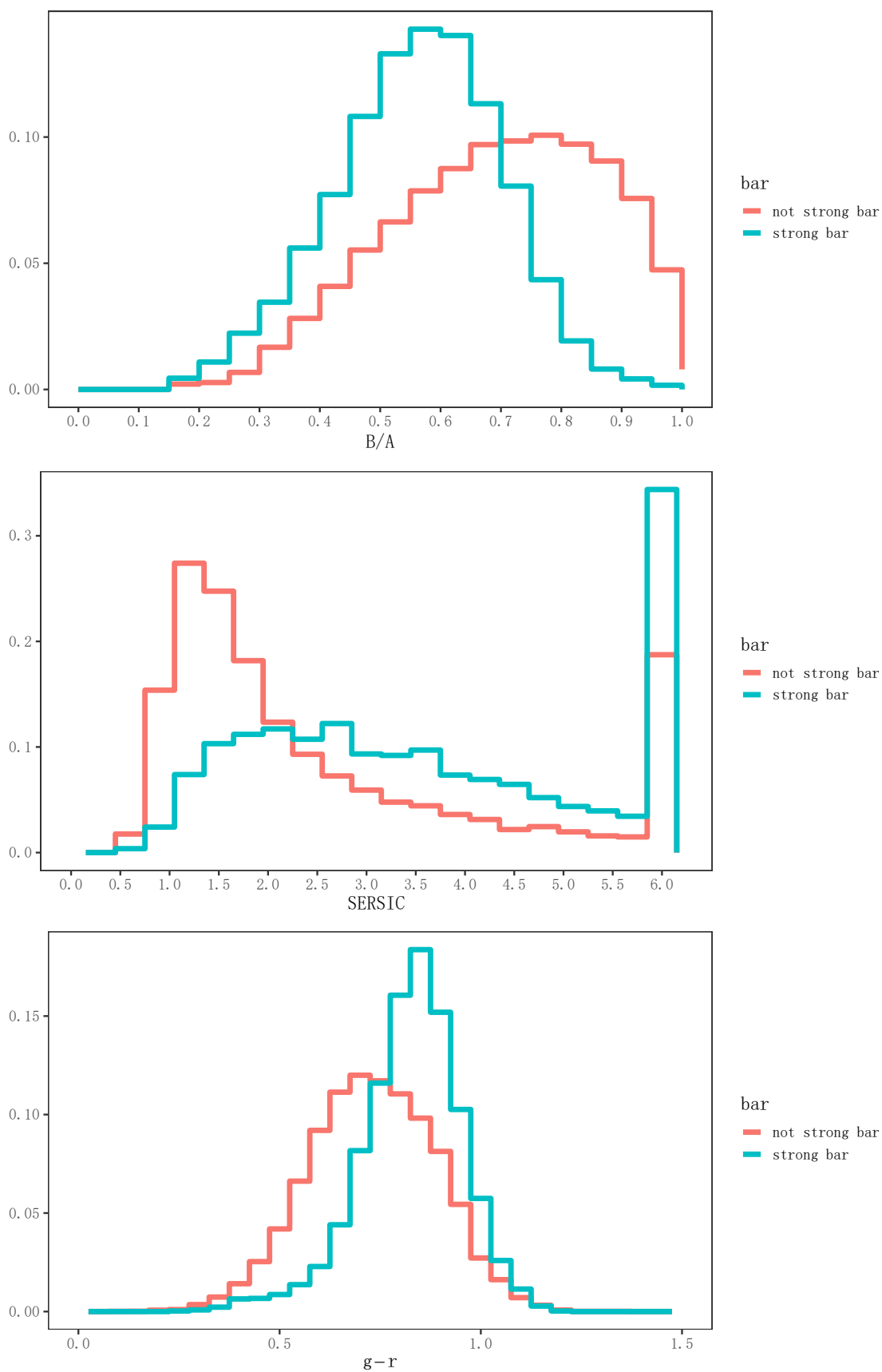


图 4-5 “强棒星系” 和 “非强棒星系” 的 b/a、SERSIC 指数、g-r 颜色的分布直方图
Figure 4-5 Histograms of the b/a, SERSIC index, and g-r color distribution for “strong bar” and “non-strong bar” galaxies

与 4.2.1 节一样，我们的测试集样本占棒星系样本的 30%，样本数为 7951 个。其中“有棒星系”和“无棒星系”分别为 3301 个和 4650 个。将训练集放入模型进行训练后，我们用测试集样本放入随机森林模型进行测试。测试样本中得到的混淆矩阵如图 4-6 所示。根据混淆矩阵得到的准确率为：0.677。同时将每个特征值的 Gini 系数进行归一化并输出，视为每个特征与棒的相关性，如图 4-7 所示。

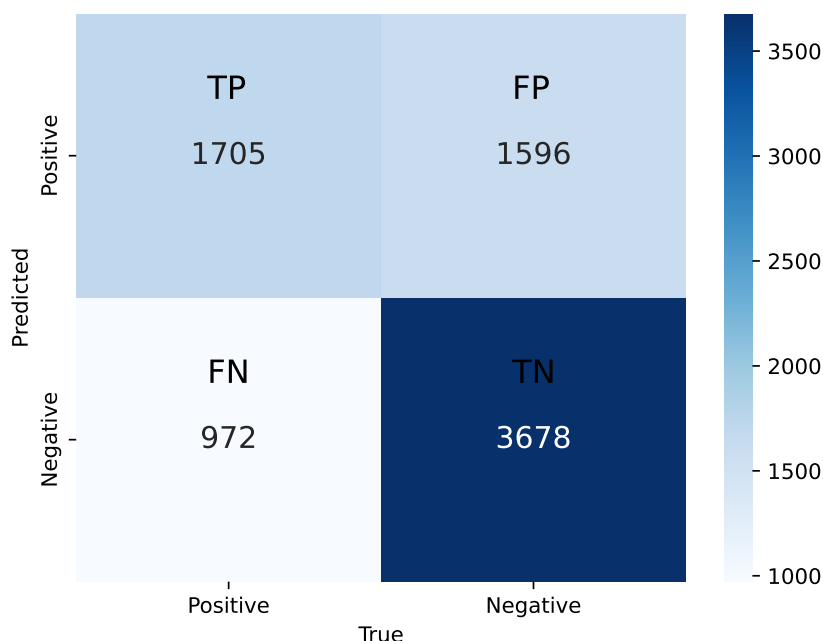


图 4-6 “有棒星系”和“无棒星系”中的随机森林模型中的混淆矩阵
 Figure 4-6 The confusion matrix of the random forest model for "have bar" and "no bar" galaxies

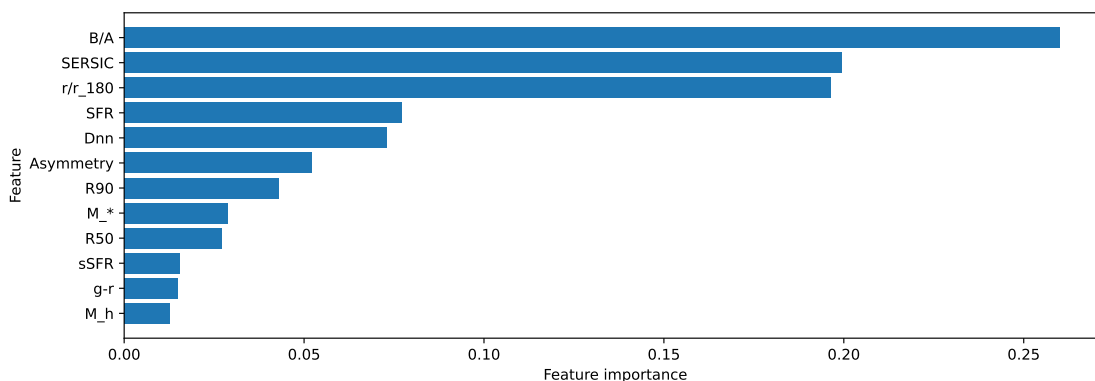


图 4-7 “有棒星系”和“无棒星系”中的随机森林模型得到的参数重要性排序
 Figure 4-7 The ranking of feature importance obtained from the random forest model

根据 Gini 系数可以看出，在“有棒星系”和“无棒星系”中的随机森林模型中，星系的轴比，Sersic 指数和 r/r_{180} 的重要性较高，重要性系数都高于 0.1，其

余参数重要性都较低。比较有趣的是，在有棒-无棒分类中，星系的颜色不再呈现重要的相关而是 r/r_{180} 。这一点是有点出乎我们意料的。“有棒星系”和“无棒星系”样本中星系的轴比，Sersic 指数和 r/r_{180} 这三个参数的分布直方图如图 4-8 所示。可以看到在轴比和 Sersic 指数的分布上，与“强棒星系”和“非强棒星系”模型类似，“有棒星系”样本在 b/a 上的分布更偏向于偏小的一端，而“无棒星系”偏向于 b/a 较大的一端。在 Sersic 指数的分布上，“有棒星系”样本更偏向于较大的 Sersic 指数，“无棒星系”样本更偏向于较小的 Sersic 指数。而在 r/r_{180} 分布上，“有棒星系”样本偏向于更大的 r/r_{180} 上，而“无棒星系”则相反。这一点非常有趣，甚至有点出人意料。其可能隐含了星系在初次内落进入更大质量暗晕的阶段可以触发弱棒特征的物理图像。作为这个推论的旁证之一就是在第三章中，我们从图 3.6 可以看到弱棒星系具有最高的比例处于 Infalling 区域。显然，该现象值得在今后的工作中进一步深入探讨。

同时根据模型的准确率来看，模型整体预言没有“强棒星系”和“非强棒星系”好，这说明强棒相对于弱棒来说是星系的一种特殊状态，而无棒和弱棒星系之间的差异并不十分显著。这和第 3 章得到的结论是一致的

其余参数重要性的排列如下：SFR、Dnn、Asymmetry、 R_{90} 、 M_* 、 R_{50} 、sSFR、g-r 和 M_h 。但是，整体贡献的相关性都小于 0.1。这些弱相关性可能涉及非常复杂的物理过程，本文就不再一一讨论。

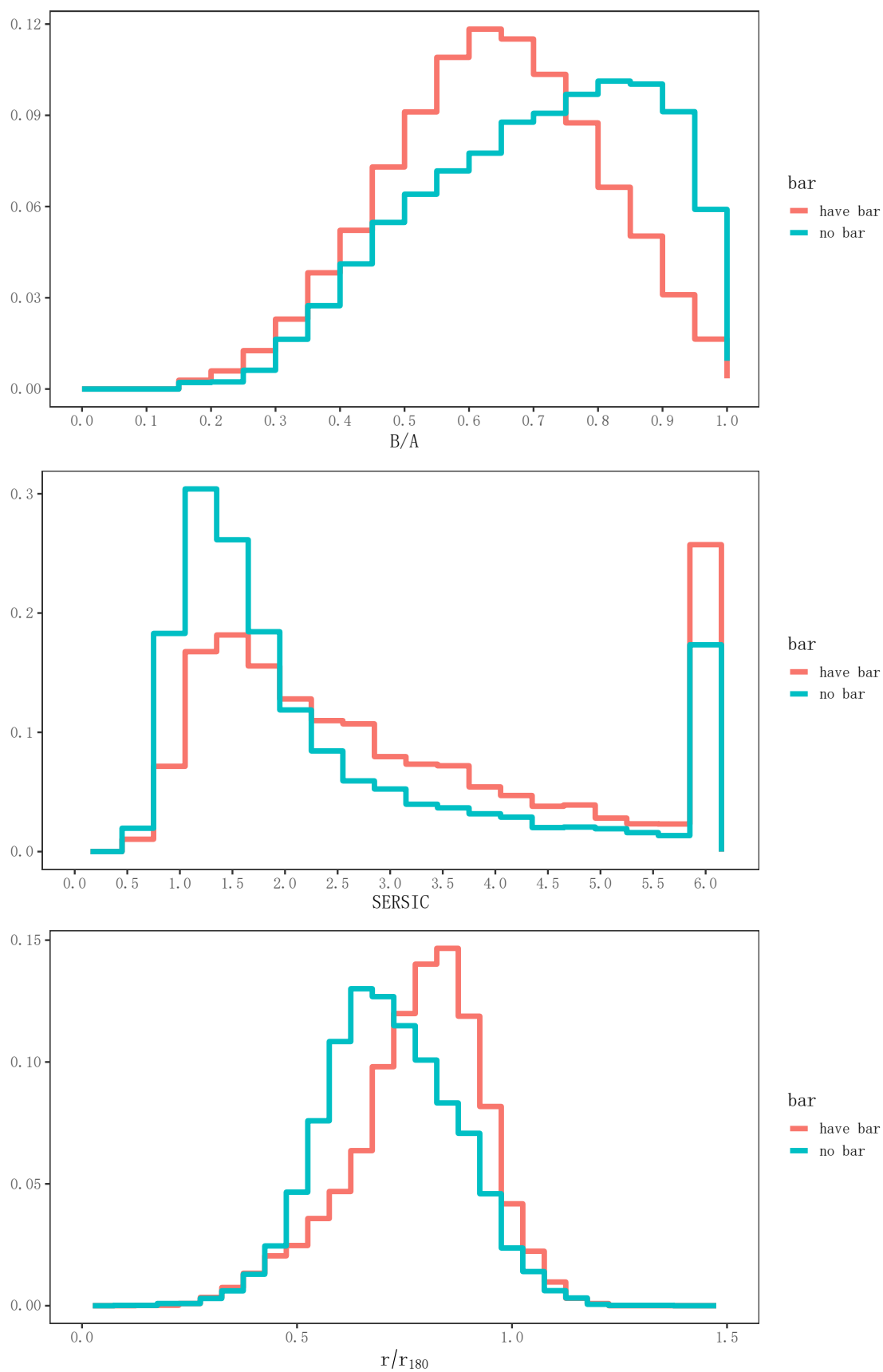


图 4-8 “有棒星系” 和 “无棒星系” 的 b/a 、SERSIC 指数、 r/r_{180} 的分布直方图
 Figure 4-8 Histograms of the b/a , SERSIC index, and r/r_{180} distribution for “have bar” and “no bar” galaxies

第 5 章 总结与展望

在本工作中，我们选取了 Galaxy Zoo DECaLS 数据中的棒旋星系样本，结合 GALEX-SDSS-WISE Legacy Catalogue-X2(GSWLC-X2) 数据中的恒星形成信息，Yang et al. (2007) 星系群样本的星系所处环境信息，以及 NSA 数据中的星系形态参数信息，得到了目前为止最大的强棒和弱棒星系样本。我们首先对比分析了不同棒特征星系的在质量、恒星形成和环境特征等物理参量的分布上存在的差别。在通过构建控制样本控制恒星质量和环境参量之后，我们系统的对棒旋星系中的棒与星系自身的恒星形成之间的相关性进行了统计分析。并训练随机森林模型，来对星系中各个物理参数与棒的相关性进行排序。具体结论如下：

1. 有棒星系相较于无棒星系，拥有更大的恒星质量和更低的 sSFR，处于更大的暗晕，更多的处在位力化的区域。相关物理特征的差别，强棒比弱棒更为显著。在相同的质量和环境影响下，棒旋星系的“熄灭”比例更多。同样，该效应中，强棒星系比弱棒更为显著。在相同的质量和环境影响下，对于仍处于恒星形成主序区域内的星系的来说，有棒和无棒星系的整体 sSFR 并无明显区别。我们的研究首次系统性的分离了环境因素可能对星系 sSFR 带来的偏差，在此基础上得到的有棒和无棒星系的 sSFR 特征的差别更能反映星系的棒特征和恒星形成性质的物理上的关联，对于棒旋星系的研究具有参考价值。

2. 我们分别训练了两个随机森林模型，区别只在于弱棒星系的归类，一种将其归为强棒星系一类，另一种将其归为无棒星系一类，得到的预测准确率分别为 0.677 和 0.719。表明弱棒星系在星系形态上更偏向于无棒星系，而不是强棒星系。这个结论与 Kim et al. (2017) 的结论符合，在他们的工作中发现弱棒星系在比恒星形成率 (sSFR)、多个波段的颜色 (g-r, NUV-r, mid-infrared[3.4]-[12]) 和中性氢 (HI) 等参数上与无棒星系很相似。在将每个特征的重要性进行排序后，发现星系的轴比、星系 Sersic 指数，g-r 颜色和 r/r_{180} 对于模型进行棒的判据较为重要。

我们研究中的 sSFR 描述的星系整体的恒星形成特征，这一整体特征可能对星系中的局域恒星形成过程并不敏感。而棒特征对星系恒星形成特征的影响可能并不是整体的，而是局部的。比如，Lin et al. (2017) 的研究表明，棒特征仅对星系中心区域的 SFR 存在一定抑制作用。如果星系中心部分的 SFR 占据星系整体 SFR 的比例不大的话，那么中心部分 SFR 的抑制并不能显著降低星系整体的 SFR。这样，我们就不一定能探测到有棒的恒星形成星系的平均 sSFR 的显著降低。要进一步探讨星系中的棒特征对恒星形成影响的细节物理过程，我们可以从两方面进行拓展。一方面，我们可以细化星系中棒的物理特征，比如定量计算棒的强度 (Géron et al., 2021) 而不是采用本文中的人眼得到的简单的强弱特征；另一方面，我们可以参考 Lin et al. (2017) 的研究工作，在大样本的基础上，借助多波段图像进行不同区域的 SED 拟合，从而细致分析棒特征对星系内部物理恒星

形成过程的影响。同时本文只使用了一个颜色，不能完全反映星系的光谱能量分布特征，未来工作还可以引入更多颜色特征如 NUV-r 等。

我们在随机森林模型中发现的与棒特征存在相关性的物理参数及重要性排序对研究棒的起源给出了一些线索。但是，相关性不等于因果性，这些因素有些可能是棒特征产生的因，有些是棒特征导致的果。要区分因果关系，还需要结合星系棒的形成和演化的理论模型及数值模拟等结果 (如 [Rosas-Guevara et al. \(2022\)](#); [Bi et al. \(2022\)](#)) 进行进一步的分析和讨论。

中国空间站巡天空间望远镜项目 (CSST) 是我国即将投入运行的重大科学项目，其口径为 2 米，巡天视场面积 ≥ 1.1 平方度，空间分辨本领好于 $0.15''$ 。同时巡天成像观测波段覆盖 $0.255 \sim 1.0 \mu\text{m}$ ，可以得到覆盖近紫外、光学、红外等多个波段 (NUV、u、g、r、i、z、y) 的数据。在目前的成像任务规划中，包括 17500 平方度多色成像观测和 400 平方度的多色成像深场观测，多色成像观测中不同波段的极限星等从 24.4 等到 26.3 等，多色成像深场观测中不同波段的极限星等从 25.7 等到 27.5 等。故 CSST 能进行更深的测光巡天，得到更高分辨率的图像，能得到更多的棒旋星系样本，可帮助进行更细致的棒旋星系物理性质分析。

参考文献

- Aguerri J A L, Méndez-Abreu J, Corsini E M, 2009. The population of barred galaxies in the local universe. I. Detection and characterisation of bars[J/OL]. , 495(2): 491-504. DOI: [10.1051/0004-6361:200810931](https://doi.org/10.1051/0004-6361:200810931).
- Athanassoula E, 2003. What determines the strength and the slowdown rate of bars?[J/OL]. , 341(4): 1179-1198. DOI: [10.1046/j.1365-8711.2003.06473.x](https://doi.org/10.1046/j.1365-8711.2003.06473.x).
- Ayyar V, Knop R, Awbrey A, et al., 2022. Identifying Transient Candidates in the Dark Energy Survey Using Convolutional Neural Networks[J/OL]. , 134(1039): 094501. DOI: [10.1088/1538-3873/ac8375](https://doi.org/10.1088/1538-3873/ac8375).
- Barazza F D, Jablonka P, Desai V, et al., 2009. Frequency and properties of bars in cluster and field galaxies at intermediate redshifts[J/OL]. , 497(3): 713-728. DOI: [10.1051/0004-6361/200810352](https://doi.org/10.1051/0004-6361/200810352).
- Beers T C, Flynn K, Gebhardt K, 1990. Measures of Location and Scale for Velocities in Clusters of Galaxies—A Robust Approach[J/OL]. , 100: 32. DOI: [10.1086/115487](https://doi.org/10.1086/115487).
- Bi D, Shlosman I, Romano-Díaz E, 2022. Modeling Evolution of Galactic Bars at Cosmic Dawn [J/OL]. , 934(1): 52. DOI: [10.3847/1538-4357/ac779b](https://doi.org/10.3847/1538-4357/ac779b).
- Boselli A, Fossati M, Sun M, 2022. Ram pressure stripping in high-density environments[J/OL]. , 30(1): 3. DOI: [10.1007/s00159-022-00140-3](https://doi.org/10.1007/s00159-022-00140-3).
- Bureau M, Freeman K C, 1999. The Nature of Boxy/Peanut-Shaped Bulges in Spiral Galaxies[J/OL]. , 118(1): 126-138. DOI: [10.1086/300922](https://doi.org/10.1086/300922).
- Burhanudin U F, Maund J R, Killestein T, et al., 2021. Light-curve classification with recurrent neural networks for GOTO: dealing with imbalanced data[J/OL]. , 505(3): 4345-4361. DOI: [10.1093/mnras/stab1545](https://doi.org/10.1093/mnras/stab1545).
- Cavanagh M K, Bekki K, Groves B A, et al., 2022. The evolution of barred galaxies in the EAGLE simulations[J/OL]. , 510(4): 5164-5178. DOI: [10.1093/mnras/stab3786](https://doi.org/10.1093/mnras/stab3786).
- Cervantes Sodi B, 2017. Stellar, Gas, and Dark Matter Content of Barred Galaxies[J/OL]. , 835(1): 80. DOI: [10.3847/1538-4357/835/1/80](https://doi.org/10.3847/1538-4357/835/1/80).
- Cervantes Sodi B, Li C, Park C, 2015. Dark Matter Halos of Barred Disk Galaxies[J/OL]. , 807(1): 111. DOI: [10.1088/0004-637X/807/1/111](https://doi.org/10.1088/0004-637X/807/1/111).
- Chabrier G, 2003. Galactic Stellar and Substellar Initial Mass Function[J/OL]. , 115(809): 763-795. DOI: [10.1086/376392](https://doi.org/10.1086/376392).
- Chi H, Wei S, Wang F, et al., 2023. Identifying 46 New Open Cluster Candidates in Gaia EDR3 Using a Hybrid pyUPMASK and Random Forest Method[J/OL]. , 265(1): 20. DOI: [10.3847/1538-4365/acb2cc](https://doi.org/10.3847/1538-4365/acb2cc).
- Consolandi G, Dotti M, Boselli A, et al., 2017. Bars as seen by Herschel and Sloan[J/OL]. , 598: A114. DOI: [10.1051/0004-6361/201629213](https://doi.org/10.1051/0004-6361/201629213).
- Curti M, Hayden-Pawson C, Maiolino R, et al., 2022. What drives the scatter of local star-forming galaxies in the BPT diagrams? A Machine Learning based analysis[J/OL]. , 512(3): 4136-4163. DOI: [10.1093/mnras/stac544](https://doi.org/10.1093/mnras/stac544).
- de Souza R S, Quanfeng X, Shen S, et al., 2022. qrpca: A package for fast principal component analysis with GPU acceleration[J/OL]. *Astronomy and Computing*, 41: 100633. DOI: [10.1016/j.ascom.2022.100633](https://doi.org/10.1016/j.ascom.2022.100633).

- Dey A, Schlegel D J, Lang D, et al., 2019. Overview of the DESI Legacy Imaging Surveys[J/OL]. , 157(5): 168. DOI: [10.3847/1538-3881/ab089d](https://doi.org/10.3847/1538-3881/ab089d).
- Donnari M, Pillepich A, Nelson D, et al., 2021. Quenched fractions in the IllustrisTNG simulations: comparison with observations and other theoretical models[J/OL]. , 506(4): 4760-4780. DOI: [10.1093/mnras/stab1950](https://doi.org/10.1093/mnras/stab1950).
- Erwin P, 2018. The dependence of bar frequency on galaxy mass, colour, and gas content - and angular resolution - in the local universe[J/OL]. , 474(4): 5372-5392. DOI: [10.1093/mnras/stx3117](https://doi.org/10.1093/mnras/stx3117).
- Erwin P, Debattista V P, 2017. The frequency and stellar-mass dependence of boxy/peanut-shaped bulges in barred galaxies[J/OL]. , 468(2): 2058-2080. DOI: [10.1093/mnras/stx620](https://doi.org/10.1093/mnras/stx620).
- Fang G, Ba S, Gu Y, et al., 2023. Automatic Classification of Galaxy Morphology: A Rotationally-invariant Supervised Machine-learning Method Based on the Unsupervised Machine-learning Data Set[J/OL]. , 165(2): 35. DOI: [10.3847/1538-3881/aca1a6](https://doi.org/10.3847/1538-3881/aca1a6).
- Flaugher B, Diehl H T, Honscheid K, et al., 2015. The Dark Energy Camera[J/OL]. , 150(5): 150. DOI: [10.1088/0004-6256/150/5/150](https://doi.org/10.1088/0004-6256/150/5/150).
- Font J, Beckman J E, Martínez-Valpuesta I, et al., 2017. Kinematic Clues to Bar Evolution for Galaxies in the Local Universe: Why the Fastest Rotating Bars are Rotating Most Slowly[J/OL]. , 835(2): 279. DOI: [10.3847/1538-4357/835/2/279](https://doi.org/10.3847/1538-4357/835/2/279).
- Géron T, Smethurst R J, Lintott C, et al., 2021. Galaxy zoo: stronger bars facilitate quenching in star-forming galaxies[J/OL]. , 507(3): 4389-4408. DOI: [10.1093/mnras/stab2064](https://doi.org/10.1093/mnras/stab2064).
- Ghosh S, Saha K, Di Matteo P, et al., 2021. Fate of stellar bars in minor merger of galaxies[J/OL]. , 502(2): 3085-3100. DOI: [10.1093/mnras/stab238](https://doi.org/10.1093/mnras/stab238).
- Guo Y, Jogee S, Finkelstein S L, et al., 2023. First Look at $z > 1$ Bars in the Rest-frame Near-infrared with JWST Early CEERS Imaging[J/OL]. , 945(1): L10. DOI: [10.3847/2041-8213/acacfb](https://doi.org/10.3847/2041-8213/acacfb).
- Huertas-Company M, Rouan D, Tasca L, et al., 2008. A robust morphological classification of high-redshift galaxies using support vector machines on seeing limited images. I. Method description [J/OL]. , 478(3): 971-980. DOI: [10.1051/0004-6361:20078625](https://doi.org/10.1051/0004-6361:20078625).
- Jaffé Y L, Smith R, Candlish G N, et al., 2015. BUDHIES II: a phase-space view of H I gas stripping and star formation quenching in cluster galaxies[J/OL]. , 448(2): 1715-1728. DOI: [10.1093/mnras/stv100](https://doi.org/10.1093/mnras/stv100).
- Kim E, Hwang H S, Chung H, et al., 2017. Star Formation Activity of Barred Spiral Galaxies[J/OL]. , 845(2): 93. DOI: [10.3847/1538-4357/aa80db](https://doi.org/10.3847/1538-4357/aa80db).
- Knapen J H, Beckman J E, Shlosman I, et al., 1995. The Striking Near-Infrared Morphology of the Inner Region in M100[J/OL]. , 443: L73. DOI: [10.1086/187839](https://doi.org/10.1086/187839).
- Kormendy J, Kennicutt J, Robert C., 2004. Secular Evolution and the Formation of Pseudobulges in Disk Galaxies[J/OL]. , 42(1): 603-683. DOI: [10.1146/annurev.astro.42.053102.134024](https://doi.org/10.1146/annurev.astro.42.053102.134024).
- Kruk S J, Lintott C J, Bamford S P, et al., 2018. Galaxy Zoo: secular evolution of barred galaxies from structural decomposition of multiband images[J/OL]. , 473(4): 4731-4753. DOI: [10.1093/mnras/stx2605](https://doi.org/10.1093/mnras/stx2605).
- Lee G H, Park C, Lee M G, et al., 2012. Dependence of Barred Galaxy Fraction on Galaxy Properties and Environment[J/OL]. , 745(2): 125. DOI: [10.1088/0004-637X/745/2/125](https://doi.org/10.1088/0004-637X/745/2/125).
- Lin L, Li C, He Y, et al., 2017. Bar-induced Central Star Formation as Revealed by Integral Field Spectroscopy from CALIFA[J/OL]. , 838(2): 105. DOI: [10.3847/1538-4357/aa657a](https://doi.org/10.3847/1538-4357/aa657a).
- Łokas E L, 2020. An interesting case of the formation and evolution of a barred galaxy in the cosmological context[J/OL]. , 642: L12. DOI: [10.1051/0004-6361/202039425](https://doi.org/10.1051/0004-6361/202039425).

- Łokas E L, Ebrova I, del Pino A, et al., 2016. Tidally Induced Bars of Galaxies in Clusters[J/OL]. , 826(2): 227. DOI: [10.3847/0004-637X/826/2/227](https://doi.org/10.3847/0004-637X/826/2/227).
- Marinova I, Jogee S, Heiderman A, et al., 2009. Barred Galaxies in the Abell 901/2 Supercluster with Stages[J/OL]. , 698(2): 1639-1658. DOI: [10.1088/0004-637X/698/2/1639](https://doi.org/10.1088/0004-637X/698/2/1639).
- Martinez-Valpuesta I, Aguerri J A L, Gonzalez-Garca A C, et al., 2017. A numerical study of interactions and stellar bars[J/OL]. , 464(2): 1502-1511. DOI: [10.1093/mnras/stw2500](https://doi.org/10.1093/mnras/stw2500).
- Masters K L, Nichol R C, Haynes M P, et al., 2012. Galaxy Zoo and ALFALFA: atomic gas and the regulation of star formation in barred disc galaxies[J/OL]. , 424(3): 2180-2192. DOI: [10.1111/j.1365-2966.2012.21377.x](https://doi.org/10.1111/j.1365-2966.2012.21377.x).
- Merrifield M R, Kuijken K, 1999. Hidden bars and boxy bulges[J/OL]. , 345: L47-L50. DOI: [10.48550/arXiv.astro-ph/9904158](https://doi.org/10.48550/arXiv.astro-ph/9904158).
- Mong Y L, Ackley K, Killestein T L, et al., 2023. Self-supervised clustering on image-subtracted data with deep-embedded self-organizing map[J/OL]. , 518(1): 752-762. DOI: [10.1093/mnras/stac3103](https://doi.org/10.1093/mnras/stac3103).
- Muzzin A, van der Burg R F J, McGee S L, et al., 2014. The Phase Space and Stellar Populations of Cluster Galaxies at $z \sim 1$: Simultaneous Constraints on the Location and Timescale of Satellite Quenching[J/OL]. , 796(1): 65. DOI: [10.1088/0004-637X/796/1/65](https://doi.org/10.1088/0004-637X/796/1/65).
- Nair P B, Abraham R G, 2010a. A Catalog of Detailed Visual Morphological Classifications for 14,034 Galaxies in the Sloan Digital Sky Survey[J/OL]. , 186(2): 427-456. DOI: [10.1088/0067-0049/186/2/427](https://doi.org/10.1088/0067-0049/186/2/427).
- Nair P B, Abraham R G, 2010b. On the Fraction of Barred Spiral Galaxies[J/OL]. , 714(2): L260-L264. DOI: [10.1088/2041-8205/714/2/L260](https://doi.org/10.1088/2041-8205/714/2/L260).
- Peng Y J, Lilly S J, Kovač K, et al., 2010. Mass and Environment as Drivers of Galaxy Evolution in SDSS and zCOSMOS and the Origin of the Schechter Function[J/OL]. , 721(1): 193-221. DOI: [10.1088/0004-637X/721/1/193](https://doi.org/10.1088/0004-637X/721/1/193).
- Petersson J, Renaud F, Agertz O, et al., 2023. From starburst to quenching: merger-driven evolution of the star formation regimes in a shell galaxy[J/OL]. , 518(3): 3261-3273. DOI: [10.1093/mnras/stac3136](https://doi.org/10.1093/mnras/stac3136).
- Rosas-Guevara Y, Bonoli S, Dotti M, et al., 2022. The evolution of the barred galaxy population in the TNG50 simulation[J/OL]. , 512(4): 5339-5357. DOI: [10.1093/mnras/stac816](https://doi.org/10.1093/mnras/stac816).
- Sadeh I, Abdalla F B, Lahav O, 2016. ANNz2: Photometric Redshift and Probability Distribution Function Estimation using Machine Learning[J/OL]. , 128(968): 104502. DOI: [10.1088/1538-3873/128/968/104502](https://doi.org/10.1088/1538-3873/128/968/104502).
- Salim S, 2014. Green Valley Galaxies[J/OL]. Serbian Astronomical Journal, 189: 1-14. DOI: [10.2298/SAJ1489001S](https://doi.org/10.2298/SAJ1489001S).
- Salim S, Lee J C, Janowiecki S, et al., 2016. GALEX-SDSS-WISE Legacy Catalog (GSWLC): Star Formation Rates, Stellar Masses, and Dust Attenuations of 700,000 Low-redshift Galaxies[J/OL]. , 227(1): 2. DOI: [10.3847/0067-0049/227/1/2](https://doi.org/10.3847/0067-0049/227/1/2).
- Salim S, Boquien M, Lee J C, 2018. Dust Attenuation Curves in the Local Universe: Demographics and New Laws for Star-forming Galaxies and High-redshift Analogs[J/OL]. , 859(1): 11. DOI: [10.3847/1538-4357/aabf3c](https://doi.org/10.3847/1538-4357/aabf3c).
- Sarkar S, Pandey B, Bhattacharjee S, 2021. Do galactic bars depend on environment?: an information theoretic analysis of Galaxy Zoo 2[J/OL]. , 501(1): 994-1001. DOI: [10.1093/mnras/staa3665](https://doi.org/10.1093/mnras/staa3665).
- Sellwood J A, 2014. Secular evolution in disk galaxies[J/OL]. Reviews of Modern Physics, 86(1): 1-46. DOI: [10.1103/RevModPhys.86.1](https://doi.org/10.1103/RevModPhys.86.1).

- Tahmasebzadeh B, Zhu L, Shen J, et al., 2022. Orbit-superposition Dynamical Modeling of Barred Galaxies[J/OL]. , 941(2): 109. DOI: [10.3847/1538-4357/ac9df6](https://doi.org/10.3847/1538-4357/ac9df6).
- Tawfeek A A, Cervantes Sodi B, Fritz J, et al., 2022. Morphology Driven Evolution of Barred Galaxies in OMEGAWINGS Clusters[J/OL]. , 940(1): 1. DOI: [10.3847/1538-4357/ac9976](https://doi.org/10.3847/1538-4357/ac9976).
- Teimoorinia H, Bluck A F L, Ellison S L, 2016. An artificial neural network approach for ranking quenching parameters in central galaxies[J/OL]. , 457(2): 2086-2106. DOI: [10.1093/mnras/stw036](https://doi.org/10.1093/mnras/stw036).
- van den Bergh S, 2002. Bar Galaxies and Their Environments[J/OL]. , 124(2): 782-785. DOI: [10.1086/341647](https://doi.org/10.1086/341647).
- Walmsley M, Lintott C, Géron T, et al., 2022. Galaxy Zoo DECaLS: Detailed visual morphology measurements from volunteers and deep learning for 314 000 galaxies[J/OL]. , 509(3): 3966-3988. DOI: [10.1093/mnras/stab2093](https://doi.org/10.1093/mnras/stab2093).
- Yang X, Mo H J, van den Bosch F C, et al., 2007. Galaxy Groups in the SDSS DR4. I. The Catalog and Basic Properties[J/OL]. , 671(1): 153-170. DOI: [10.1086/522027](https://doi.org/10.1086/522027).
- Zana T, Dotti M, Capelo P R, et al., 2018. Bar resilience to flybys in a cosmological framework [J/OL]. , 479(4): 5214-5219. DOI: [10.1093/mnras/sty1850](https://doi.org/10.1093/mnras/sty1850).
- Zana T, Lupi A, Bonetti M, et al., 2022. Morphological decomposition of TNG50 galaxies: methodology and catalogue[J/OL]. , 515(1): 1524-1543. DOI: [10.1093/mnras/stac1708](https://doi.org/10.1093/mnras/stac1708).

致 谢

在这里我要感谢西华师大的李静老师，没有您的帮助我可能就无法开启天文的道路；感谢星团组的老师们，你们的工作态度和工作氛围让我受益匪浅；感谢星团组、KKC 组内的师兄师姐师弟师妹们的帮助，希望你们能享受科研；感谢研究生办的储怡老师、马金老师，对我台里的工作生活中给了很多帮助；感谢潘翠云师姐在我答辩时期给与的帮助；感谢 1712 办公室提供的温暖和欢乐；感谢顾敏峰老师、李双良老师以及台羽协成员在我担任组织工作时提供的支持，希望你们能一直享受羽毛球；感谢我的家人对我一直以来的支持与鼓励，这对我来说是最重要的。

同时，我要感谢沈世银老师、Rafael 老师和 Ana 老师。沈老师对于问题的思考方式，对于工作的理解和态度，都让我学到了很多。尽管有时并不理解，但总能在事后的某些时候恍然大悟。对于给您这些年造成的烦恼和无奈我感到由衷的歉意。Rafael 老师和 Ana 老师对我来说不仅是工作上的标杆也是生活中难得的朋友，希望以后能在上海或是别的国家再次碰面，继续潇洒。

最后，我要感谢我的女朋友箫箫，在我读研最后阶段比较痛苦比较彷徨的时候，是你陪伴着我，安抚我的心态，与我一起理性分析面临的困境，你对我的学业起到了不可替代的作用。在即将开始的人生新篇章，希望我们相互支持继续走下去。

2023 年 6 月

作者简介及攻读学位期间发表的学术论文与其他相关学术成果

作者简介：

2014年09月——2018年06月，在河北农业大学理学院获得学士学位。

2019年09月——2023年06月，在中国科学院上海天文台攻读硕士学位。

已发表（或正式接受）的学术论文：

- (1) 棒与星系整体恒星形成性质关系的统计研究，天文学进展，2023

