北京大学

本科生毕业论文

题目：

银纬正负五度内银道面背景类星体的选源

Quasars Selection Behind the Galactic Plane within |b|<5 deg

姓　　名：刘行健

学　　号：1900011631

院　　系：物理学院

专　　业：天文学

导师姓名：吴学兵

二〇二三年 六 月

版权声明

摘要

银道面背景类星体是研究银盘气体结构分布的重要工具，也是重要的天体测量参考源，有助于提高巡天中天体测量数据的精度。然而由于低银纬消光严重、天体污染源多，对银道面背景类星体的选源比高银纬类星体的选源更加困难。扩充低银纬的类星体样本可以帮助巡天任务将天体测量的区域拓展到银道面中心，以便我们进一步研究银道面的性质。本文基于 Fu et al. (2021)构建的迁移学习银道面背景类星体选源方法——通过模拟样本和 XGBoost 两步二分类的方法减轻数据集偏移问题，并在多个方面进行改进后，将其应用到探测难度更大的银纬$|b| < 5$ deg 区域，最终在目标天区选出 4222 个类星体候选体。我们使用有放回随机抽样对高银纬样本进行多轮模拟，使得高银纬样本的利用率分别提高到 95%和 99%以上；我们去掉了消光严重的 PS1 蓝端的 g 和 r 波段，加入了近红外 UKIDSS J,H,K 波段以提高对高红移类星体的筛选效率；我们调高 Healpixmap 的分辨率来平滑先验分布，减轻边界异常值；我们使用后验概率校正和颜色截断等手段来提高候选体的纯净度。以上改进后的类星体选源方法有助于提高密集星场中类星体的筛选效率，对未来的银道面类星体巡天具有重要的科学价值。

关键词：类星体，银道面，迁移学习，机器学习分类

ABSTRACT

Quasars behind the Galactic plane (GPQs) are useful tools for studying the gas structure of Milky Way, and provide important astrometric references that help to improve the accuracy of astrometric data. However, due to the severe extinction and high source densities in the Galactic plane, search for GPQs has long been a 'zone of avoidance' for extragalactic astronomy. Expanding the GPQ samples can help sky surveys extend their probing regions to the Galactic Center, which can ensure further investigation of the nature of Galactic Plane. This paper builds on the Transfer Learning Quasar selection method constructed by Fu et al. (2021), which applies the sample mocking method and designs a two-step binary XGBoost classification algorithm to mitigate the dataset shift caused by different probability distributions between high-b Quasars and low-b Quasars, and we improve this method in several aspects: We build a Bootstrap Mocking method and increase the selection rate of high-b QSO and Galaxy samples to 95% and 99% respectively. We remove the highly-extincted PS1 g and r bands and add UKIDSS J,H,K bands photometry to the XGBoost classification features so as to improve the selection efficiency of high redshifted Quasars. We increase the pixel numbers of Healpixmap and correct the boundary outliers to smooth the prior probability distribution of Mock GPQ samples. We also apply a further color cut to remove stellar contaminants. Results give 4222 Quasar candidates within $|b| < 5$ deg, and this modified method is of great scientific value in boosting the Quasar selection efficiency behind dense stellar fields.

KEY WORDS: Quasars, Galactic Plane, Transfer Learning, Machine Learning Classification

目录

第一章 引言

1.1 活动星系核和类星体

活动星系（Active Galaxy）相对于其他星系，其内部存在着更剧烈的物理过程，如超过恒星内部核反应的产能、高能 X 和 γ 射线、物质的喷射和爆发现象等，且在各波段的能量辐射十分巨大。我们通常根据以下几个主要观测特征来判断星系是否是 AGN：

- AGN 具有明亮的致密核区。活动星系核的光度很高，数量级在10^{43}～$10^{48} erg\ s^{-1}$，但发光区域的尺度很小，在 0.1pc 以内；
- 在许多电磁波段存在非热致连续辐射。在射电、光学、X 射线等波段，AGN 的光谱为幂律谱，且辐射是偏振的；而由于恒星的光谱为热致黑体谱，其他星系的光谱主要是该星系中恒星光谱的总和，所以其他星系的光谱主要是黑体谱。
- 存在强而宽的原子、离子发射线，发射高能光子(如 X 和 γ 射线)的能力也更强；
- 连续辐射和发射线的强度、轮廓和偏振等会随时变化。

活动星系核主要包括类星体（Quasar，简称 QSO），赛弗特星系（Seyfert Galaxies），射电星系（Radio Galaxies），蝎虎座 BL 型天体（BL Lac Objects），光学激变天体（Optically Violent Variable）等。

相比于其他活动星系核，类星体具有以下几点基本观测特征：

- 点源形态（类星）

从地面望远镜看，绝大多数类星体呈恒星状，角直径小于 1”。利用这个特征可以区分类星体和星系等有视面天体。除此之外，有些类星体可以看到喷流（通常在射电波段，或光学、X 射线波段）

- 高红移

河外天体的光谱都有红移，但其中类星体红移最大，从 0.1 到 7.6 不等。

- 高光度

类星体是最明亮的一类活动星系核，光度～$10^{42}-10^{48} erg\ s^{-1}$。

- 光变

类星体在光学波段的辐射常有变化，这是类星体的普遍特征。类星体的光变没有明显的周期性，光变时标一般为几年，光变幅度一般在 0.1-0.2 等以下。除此之外，在射电波段和 X 射线波段也常能观测到辐射变化。

- 连续光谱

类星体的连续辐射谱和恒星的黑体谱完全不同。不同波段的类星体的光谱虽然都近似遵从幂律谱，但不同波段的谱指数不尽相同，不能用单一的幂律谱来描述。这也说明各波段虽然都是非热致辐射为主，但辐射起源并不一定相同。

- 强发射线

类星体发射谱中常见的谱线有Ly α，H_β，He，Fe II, Mg II, C IV 等允许线和[O III], [O II], [N II]等禁线。发射线是区别类星体和恒星、正常星系的重要观测特征。

关于类星体的结构分布，广泛接受并使用的是活动星系核的统一模型(Antonucci,1993)，认为类星体从内到外分别为核心的超大质量黑洞（Supermassive Black Hole，简称 SMBH）、吸积盘（Accretion Disk）、冕区（Corona）、喷流（Jet）、宽线区（Broad Line Region，简称 BLR）、尘埃环（Dust Torus）、窄线区（Narrow Line Region，简称 NLR）。具体如图 1.1：

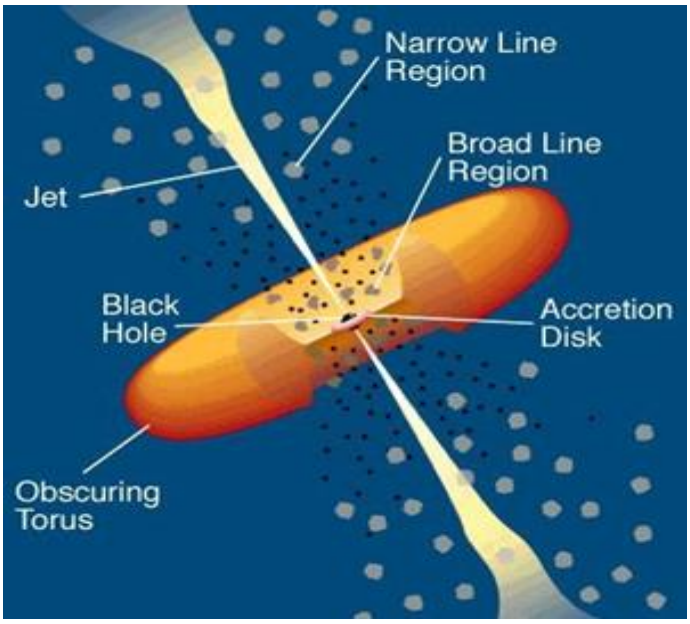

图 1.1 活动星系核模型统一模型。图源: C.M. Urry and P. Padovani

在统一模型中，活动星系核中心的超大质量黑洞通过强大的引力作用，吸积其附近的尘埃和气体等，形成一个高速旋转的巨大吸积盘。吸积盘外层存在快速转动的气体，通过粘滞作用向外传递角动量，释放引力能，产生巨大的物质喷流，而强大的磁场又约束着物质喷流只能够沿着磁轴的方向，通常是与吸积盘平面相垂直的方向高速喷出。如果这些喷流刚好对着观察者，就观测到了射电强的类星体。

类星体具有着覆盖从射电波段到 γ 波段的辐射，以及一系列强发射线，而不同波段的辐射来源于不同的区域。在吸积盘内侧靠近中心超大质量黑洞的区域温度最高，主要为 X 射线和紫外辐射；在吸积盘外侧的稀薄气体区域温度较低，主要为光学-紫外波段的连续谱辐射；吸积盘外高速转动的气体区域为宽发射线区；宽发射线区域外的冷尘埃环处为红外连续谱；吸积盘面上下分布着窄线区，比宽线区尺度更大；垂直吸积盘面的喷流向外延伸，主要为射电波段。类星体各个波段的颜色特征是我们进行选源的重要依据。

类星体的科学研究具有十分重要的天文学价值。对类星体的研究可以帮助我们了解超大质量黑洞的形成和演化、黑洞和宿主星系的共同演化过程(Di Matteo et al. 2005; Kormendy & Ho 2013)；类星体的吸收线可以用于探测宇宙不同红移处的星际介质(Trump et al. 2006)；大量活动星系核和类星体样本可以用于研究宇宙的大尺度结构及其演化过程(Eisenstein et al. 2011; Blanton et al. 2017)，测量宇宙学重子声波震荡(Zhao et al. 2018)。

1.2 类星体的选源方法

类星体的选源包括类星体候选体的选择和对候选体进行光谱证认。为了提高最终证认的类星体样本的完备性和精确度，类星体候选体的选源方法非常重要，主要是通过各种观测特征从巡天观测的天区源表内筛选出类星体候选体，然后通过利用类星体的发射线和幂律光谱等观测特征来对候选体进行观测证认，以获得最终的类星体样本。截止目前，已被类星体巡天计划观测证认的类星体有：

- The Bright Quasar Survey(BQS; Schmidt & Green, 1983)发现了 114 个类星体；
- The Large Bright Quasar Survey(LBQS; Hewett, Foltz, & Chaffee, 1995)发现了一千多个类星体；
- The 2dF Quasar Redshift Survey(2QZ; Croom et al. 2004)发现两万多个类星体；
- Sloan Digital Sky Survey 第十六次数据发布的类星体(SDSS DR16Q; Lyke et al., 2020)共有 750414 个类星体，包括 225082 个新发现的类星体。

目前为止，类星体的选源方法包括但不限于：X 射线波段选源(e.g. Grazian et al. 2000)、利用类星体的紫外超选源(ultraviolet excess; e.g. Green et al. 1986)、光学-近红外颜色选源(e.g. Wu & Jia 2010)、中红外颜色选源(e.g. Lacy et al. 2004; Mateos et al. 2012)、射电波段选源(e.g. Gregg et al. 1996; Becker et al. 2001)、利用类星体光变选源(e.g. Dobrzycki et al. 2003)等。除此之外，类星体自行为零的观测特征也是区分类星体和恒星的重要依据。

颜色选源又称色指数选源，是效率最高的选源方法之一。但是该方法有选择效应，倾向于选择颜色偏蓝的天体，而由于类星体的颜色与红移和星际消光等有关，所以该方法不能很好地筛选高红移类星体。为了扩大红移的探测范围，我们可以利用类星体的红外波段颜色特征来寻找高红移或受星际红化严重的类星体样本，如 UKIRT Infrared Deep Sky Survey（UKIDSS；Lawrence et al. 2007）的 J,H,K 波段的颜色。除此之外，利用 X 射线波段或射电波段区分恒星和类星体、利用无缝光谱寻找发射线较强的类星体都可以作为补充。可见，为了尽可能地提高样本的完备度，我们可以采用多波段选源的方法，以求各波段的探测可以互为补充。

随着机器学习在天文学领域的应用日益普及，利用机器学习进行类星体选源的方法也建立起来，如 Jin et al. (2019)利用机器学习中的 XGBoost 和 SVM 算法，从 Pan-STARRS1 (PS1; Chambers et al. 2016)、AllWISE(Wright et al. 2010; Mainzer et al. 2011)的源表中分类出类星体候选体；Bailer-Jones et al. (2019)利用高斯混合模型对 Gaia DR2(Gaia Collaboration et al. 2018b)的源表进行分类筛选出类星体，并解决了机器学习分类中类别不平衡的问题。可以预见，利用类星体的多波段特征进行机器学习分类选源可以高效地执行类星体搜寻的任务，尽可能地提高类星体候选体表的完备度。

1.3 银道面背景类星体的选源

银道面背景类星体(Galatic Plane Quasar，以下简称 GPQ)对于其所处天区的研究具有着重要的科学价值。一方面，分析银道面背景类星体的光谱吸收线可以帮助我们研究银盘气体结构(Ben Bekhti et al.2012; Westmeier 2018)；另一方面，类星体由于红移高、亮度大、自行和视差可以忽略，所以是理想的天体测量参考源。大量的银道面背景类星体样本可以帮助 Gaia 等天体测量任务建立稳定的天体测量参考架，对提高银道面天体测量数据精度具有重要的影响(Arenou et al. 2018)。由于 Gaia 在银纬$|b|<10$ deg 缺少足够的类星体参考源，尽可能地扩充更低银纬的类星体样本有助于提高低银纬天区的天体测量精度，以便我们更好地研究银道面的性质。

虽然目前为止的类星体巡天已经发现了大量的类星体，但大多数巡天计划主要把观测重点放在北天的高银纬天区，银道面天区（银纬$|b|<20$ deg）的类星体却极为稀少。在 2023 年 4 月公开的第 7.10 版百万类星体源表(Million Quasars Catalog Version 7.10, 简称 Milliquas; Flesch, 2021)中，所收集到文献认证的 I 型类星体和活动星系核有 845286 个，其中只有有

39550 个位于银纬$|b|<20$ deg，而只有 1767 个位于银纬$|b|<5$ deg。

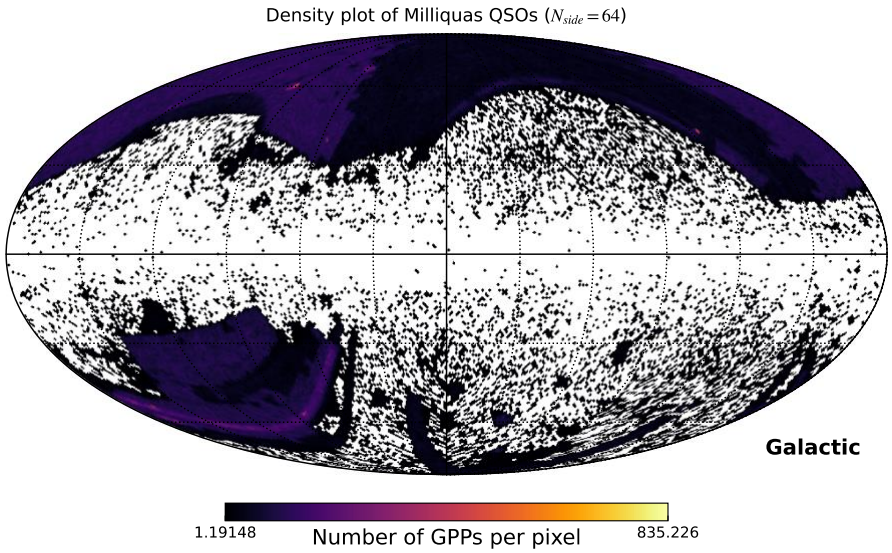

图 1.2　Milliquas 中收录的 QSO/AGN 的天区密度分布图

这是由于，相比于高银纬类星体，银道面类星体的搜寻面临着许多更严峻的困难：首先，银道面的星际消光和红化更加严重；其次，银道面天区的星场更加密集，污染严重，测光质量差；除此之外，有些天体，如白矮星、M/L/T 型矮星、年轻恒星等，与类星体的许多观测性质很相似，都是我们寻找类星体时难以剔除的污染源(e.g. Kirkpatrick et al.1997; Vennes et al. 2002; Chiu et al. 2006)。

目前学界对密集星场中类星体的搜寻已有初步的探索。如 Im et al. (2007)通过对 2MASS 巡天(Skrutskie et al. 2006))运用近红外波段 J-K>1.4 的颜色筛选确认了 40 颗银纬$|b|<20$ deg 的明亮类星体和活动星系核；Kozłowski & Kochanek (2009)将 Stern et al. (2005)中的中红外颜色筛选方法修改后在麦哲伦星云背景找到了 5000 余活动星系核；Huo et al. (2015)通过 LAMOST 在 M31 和 M33 星系附近发现了 1870 颗新类星体。

然而，高银纬和低银纬的类星体，不论是表观颜色特征还是天区的密度分布都并不相同，如果像上述研究工作一样直接照搬高银纬的选源方法应用在银道面，就会导致选源结果出现偏差，这在有监督机器学习中称为数据集偏移(Quionero-Candela et al. 2022)。为此，Fu et al. (2021)构建了一套基于迁移学习展开银道面背景类星体选源的方法，有效地减轻了机器学习中数据集偏移的问题，最终在银纬$|b|<20$ deg 的范围内找到 160946 个可靠的类星体候选体。本文基于 Fu et al. (2021)构建的选源方法，将目标天区缩小到银纬$|b|<5$ deg，以得到更加准确、完备的银道面类星体样本集。

1.4 研究方法——基于迁移学习理论的机器学习多波段选源

为了解决上述的数据集偏移问题，Fu et al. (2021)构建了一套基于迁移学习展开银道面背景类星体选源的方法，使用 SDSS DR16Q 的星表，利用 PS1 的 g,r,i,z,y 五个光学波段和 AllWISE 的 W1,W2,W3 三个中红外波段的测光数据进行 XGBoost 分类选源，最终在银纬 $|b| < 20$ deg 的范围内找到 160946 个可靠的类星体候选体。

本文基于 Fu et al. (2021)构建的选源方法，将选源的目标天区缩小到银纬$|b| < 5$ deg，改进了模拟样本的方法以扩充训练集，加入 UKIRT Infrared Deep Sky Survey（UKIDSS；Lawrence et al. 2007）近红外 J,H,K 三个波段的颜色作为分类特征，训练出 XGBoost 分类模型并应用于目标天区进行选源，再通过后验概率校正、颜色截断和自行截断的方法进一步提纯类星体候选体，以期得到完备度和纯度更高的银道面类星体候选体样本集。下面将重点介绍一下本文所应用的迁移学习选源方法。

迁移学习是机器学习的研究领域之一，指的是将已有问题的解决模型迁移应用到不同但相似的问题上。Fu et al. (2021)借鉴了迁移学习的概念(Pan & Yang 2009)，对高银纬的机器学习选源方法进行修正，再迁移应用到银道面天区。我们需要处理的数据集偏移问题具体包括两个方面：不同天区类星体由于不同消光导致的表观颜色特征的偏移，为协变量偏移；不同天区天体的先验密度分布不同导致的类别比例的偏移，为先验偏移。不同的偏移问题通过不同的方法来解决：

1.4.1 模拟样本，解决协变量偏移

在数据集层面，我们通过模拟样本的方法将高银纬天区的天体样本模拟到银道面作为机器学习的训练样本，以在尽可能扩充训练样本的同时缩小训练样本和实际天区测试样本的差异，减轻数据集的协变量偏移问题。模拟的基本原理和流程见 3.1.2，模拟方法的改进见 3.1.3。

1.4.2 两步二分类，解决先验偏移

在算法层面，我们通过将传统的三分类问题（恒星 vs 星系 vs 类星体）拆分为两步二分类问题，以减轻类别不平衡的先验偏移问题。分类流程见 4.3。

1.5 论文结构

本文在第二章介绍了构建数据集应用到的各种巡天数据源表，确定了各波段的测光质量筛选判据；在第三章详细介绍了模拟样本的改进方法，训练样本的构建，以及类星体先验概率计算的方法；在第四章分析了机器学习分类模型的效果；在第五章展示了分类模型在银纬|b|<5 deg 的应用，并通过后验概率的校正、颜色截断等手段对类星体候选体进行进一步提纯。

第二章 数据源表的介绍

我们从 SDSS，Gaia 中获取光谱证认的类星体、星系和恒星源表，并根据位置坐标 ra 和 dec 分别和 PS1,UKIDSS,Catwise 交叉获得测光数据、和 Gaia 交叉获得天体测量数据。

2.1 PS1 DR2 Photometry

Pan-STARRS1 Data Release 2（PS1; Chambers et al. 2016）巡天计划在五个光学波段 g,r,i,z,y 开展了一系列同步测光巡天，包括3π天球巡天和中深场巡天。在堆叠3π天球巡天中, g,r,i,z,y 波段点源的5σ区间均值极限星等为 23.3, 23.2, 23.1, 22.3, 21.4。

由于蓝端消光更大，所以我们并未使用 g,r 两个偏蓝波段的测光数据，而是选择精确度更高的 i,z,y 波段用于后续的选源过程。由于银道面天体密度很高，我们使用 PS1 MeanObject table 中的平均位置坐标 ra_mean,dec_mean 和平均点扩散函数星等 mean_psfmag，以求获得更精确的测光数据。i, z, y 波段的消光系数$R_i = 1.9468, R_z = 1.5097, R_y = 1.2245$，由 Wang & Chen (2019) 给出。关于消光系数的计算具体见 3.1.2。

为了得到质量更高的测光数据，我们对 i,z,y 三个波段的数据进行如下限制：

- 在三个波段中都能探测到：imag>0, zmag>0, ymag>0；
- 在 i 波段被显著探测到：imag_err < 0.2171（等同于 i 波段信噪比大于 5）；
- i 波段星等不过亮，防止探测饱和：imag>14;

值得一提的是，在制作训练样本时（见 3.1, 3.3）我们并没有限制 z 波段和 y 波段的信噪比，因为相比于 i 波段，这两个波段的探测误差较大，我们选择探测质量最好的波段限制信噪比可以最大限度地扩充训练样本的数量。但在制作测试样本 All（见 5.1）时，我们为了提高测试样本的纯净度和对正负五度天区天体分布的代表性，我们尽可能地加严对测试样本的筛选，同时限制 imag_err<0.2171，zmag_err<0.2171，ymag_err<0.2171。

我们后续在 Topcat 中进行天体数据源表和 PS1 DR2 数据库的交叉。

2.2 UKIDSS Photometry

UKIRT Infrared Deep Sky Survey（UKIDSS；Lawrence et al. 2007）是 UKIRT 在红外波段的深场巡天，一共包含 LAS, DXS, UDS, GPS, GCS 五个巡天项目，其中我们使用 LAS 和 GPS 两个巡天项目的测光数据用于选源。不同巡天项目覆盖的天区不同，探测的波段和深

度也不同：LAS 巡天的覆盖天区和 SDSS 相同为高银纬，一共有 YJHK 四个波段的测光数据，其中我们使用 JHK 三个波段；GPS 巡天的覆盖天区在正负五度之内，刚好是我们选源的天区，一共有 JHK 三个波段的测光数据。

LAS 的 JHK 三个波段的5σ极限星等[①]分别为 19.9，18.6，18.2；GPS 的 JHK 三个波段的5σ极限星等分别为 19.9，19.0，18.8[②]。根据 Wang & Chen (2019) 计算出的 J, H, K 波段的消光系数分别为$R_J = 0.7285,\ R_H = 0.4185,\ R_K = 0.2263$。

在构建高银纬 GoodQSO 和 GoodGal 样本（见 3.1）时，我们使用 LAS JHK 三个波段的测光数据，且为了提高数据质量的同时尽可能扩充样本数量，选择限制探测质量最高的 J 波段限制 Jmag_err<0.2171（等同于 J 波段信噪比大于 5）。

在构建正负五度恒星训练样本（见 3.3）和总测试样本 All（见 5.1）时，我们使用 GPS JHK 三个波段的测光数据。为了尽可能地扩充训练样本的数量，我们在制作恒星训练样本时限制 Jmag_err<0.2171（等同于 J 波段信噪比大于 5）。为了提高测试样本的纯净度和对正负五度天区天体分布的代表性，我们限制 Jmag_err<0.2171，Hmag_err<0.2171，Kmag_err<0.2171，并限制 mergedClass 不为 0 或-9[③]。

我们后续在 Topcat 中进行天体数据源表和 UKIDSS LAS DR9 和 UKIDSS GPS DR6 数据库的交叉。

2.3 CatWISE2020 Photometry

Wide-field Infrared Survey Explorer(WISE; Wright et al. 2010)提供了四个中红外波段 W1, W2, W3, W4 的全天测光数据。CatWISE2020(Marocco et al. 2021)选取了 WISE 和 NEOWISE(Mainzer et al. 2011)巡天数据中 3.4μm(w1)和 4.6μm(w2)波段的全天数据，其曝光和时间跨度是 AllWISE 的六倍，时间基线是 AllWISE 的 16 倍长，所以相比 AllWISE 囊括了更多暗弱的源，且能提供更为精确的运动测量。

W1 和 W2 波段的5σ极限星等分别为 19.6，19.3[④]。根据 Wang & Chen (2019) 计算出的消光系数分别为$R_{W1} = 0.1209,\ R_{W2} = 0.0806$。

在构建训练样本（见 3.1, 3.3）时，为了使得 W1,W2 两个波段都能被显著探测到，我们

① 注：J 波段的5σ区间指的是以 2 角秒孔径探测时点源的 Vega 星等的落在探测的5σ区间内。
② 注：这里 JHK 波段的星等均为 Vega 星等，后续需要转换为 AB 星等进行运算。
③ 注：mergedClass=0 表示探测为 Noise 的概率最大；mergedClass=-9 表示探测为 Saturation 的概率最大。具体解释见 http://wsa.roe.ac.uk/ukidssdr6/gloss_m.html
④ 注：这里 WISE 波段的星等均为 Vega 星等，后续需要转换为 AB 星等进行运算。

对两个波段的信噪比(snr)分别限制 snrW1pm > 5 and snrW2pm > 3；为了防止星等过亮探测饱和，我们对两个波段的星等分别限制 W1mproPM > 8 & W2mproPM > 7。注意，这里的星等和信噪比都将自行(pm)考虑进去，更加精确。

在构建总测试样本 All（见 5.1）时，为了提高测试样本的纯净度和对正负五度天区天体分布的代表性，我们还同时限制 cat_nb=1，表示拟合源轮廓所需成分的数量不大于 1，以尽可能地去除污染源。

我们后续在 Topcat 中进行天体数据源表和 CatWISE2020 数据库的交叉。

2.4 Gaia DR3

Gaia Data Release 3(Gaia Collaboration et al. 2022) 提供了约 4.7 亿颗恒星的精确位置、自行、视差等天体测量数据以及 G 宽带波段的星等。其中，我们将在 5.5 使用 Gaia 提供的自行(pm)对类星体候选体进行截断，进一步减少恒星污染，提纯类星体候选体表。

Gaia DR3 的 Golden Sample 子集（Creevey et al. 2022）依据恒星在赫罗图的分布，给出了六个具有高质量天体物理学参数的恒星子集。在构建恒星训练样本（见 3.3）时，我们从 Golden Sample 子集中选取 OBA 型星，并依据文献中的 fgkm_2 判据来自行选取 FGKM 型星，以构建高质量的恒星训练样本。除此之外，我们还使用了 Gaia DR3 的 Ultracool dwarfs 子集(Sarro et al. 2022)，在恒星训练样本中补充入 UCD 型星（Ultra-cool Dwarf，即超低温矮星）。

2.5 SDSS

2.5.1 SDSS Quasar Catalog: SDSS DR16Q

The Sloan Digital Sky Survey(SDSS; York et al. 2000)提供了北天银纬 30 度以上天区的测光和光谱数据，包括恒星、星系和类星体。SDSS 类星体源表 Data Release 16（SDSS DR16Q；Lyke et al. 2020）涵盖了 750,414 个类星体。

为了进一步限制数据的质量，我们从 SDSS DR16Q 中去除了 Flesch (2021)中已被确认为非类星体的 82 个源，并根据 Wu & Shen (2022)给出的系统性红移（以下称 Z_sys）进行以下提纯：

· SDSS DR16Q 的红移（以下称 Z_DR16Q）大于 0 且没有红移测量方法上的问题：

Z_DR16Q > 0 & (ZWARNING == 0 | ZWARNING == -1)

- Z_sys 大于 0 且噪声干扰不能太大：

$$Z\_sys > 0 \ \&\ (Z\_sys\_err \mathrel{!}= -1) \ \&\ (Z\_sys\_err \mathrel{!}= -2)$$

- Z_sys 的相对误差，以及 Z_sys 和 Z_DR16Q 的相对差异不能太大：

$$(Z\_sys\_err/(1+Z\_sys) < 0.002) \ \&\ (abs(Z\_sys-Z\_DR16Q)/(1+Z\_sys) < 0.002)$$

根据以上判据选出的 SDSS DR16Q 的源有 421,959 个，记作 SDSS_DR16Q_propclean。

2.5.2 SDSS spectroscopically identified Galaxies: SDSS DR17Gal

我们从 SDSS Data Release 17（Abdurro'uf et al. 2022）的 SpecPhotoAll 表中选取星系样本，记作 SDSS_DR17_Gal。

2.6 优质源的筛选判据

为了提高测光数据的质量，我们对 PS1, UKIDSS, CatWISE 的波段进行星等和信噪比的限制，用于构建高质量训练样本。见下表：

表 2.1 训练样本的优质源筛选判据

判据	物理意义
snrW1pm > 5 & snrW2pm > 3	限制(考虑自行的)W1 波段的信噪比
W1mproPM > 8 & W2mproPM > 7	限制(考虑自行的)W1 波段的星等
imag>14 & ymag>0 & zmag>0	限制 PS1 i,z,y 波段的星等
j_1AperMag3Err < 0.2171	限制 UKIDSS J 波段的信噪比
e_imag < 0.2171	限制 PS1 i 波段的信噪比

对于测试样本的筛选，我们则采取更加严格的判据，以提高测试样本对目标天区的代表性，保证类星体候选体的纯度。见下表：

表 2.2 测试样本的优质源筛选判据

判据	物理意义
snrW1pm > 5 & snrW2pm > 3	限制(考虑自行的)W1 波段的信噪比
W1mproPM > 8 & W2mproPM > 7	限制(考虑自行的)W1 波段的星等
imag>14 & ymag>0 & zmag>0	限制 PS1 i,z,y 波段的星等
j_1AperMag3Err < 0.2171	限制 UKIDSS J,H,K 波段的信噪比

& hAperMag3Err < 0.2171 & kAperMag3Err < 0.2171	
e_imag < 0.2171 & e_zmag < 0.2171 & e_ymag < 0.2171	限制 PS1 i,z,y 波段的信噪比
gps_mergedclass != 0 & gps_mergedclass != 9	限制 UKIDSS GPS 探测到的类别不是 0（代表噪声）或 9（代表饱和）
cat_nb == 1	限制 CatWISE2020 用来拟合该源轮廓所需要的成分数量不大于 1

第三章 训练样本的构建

3.1 类星体和星系样本的构建——样本模拟

如 1.3 中所讲，由于银道面天区已知的类星体和星系的数量太少，需要利用高银纬天区已知的类星体和星系，通过消光校正将其模拟到银道面正负五度，来构建正负五度天区的类星体、星系的训练样本，以此解决数据集偏移中的协变量偏移问题。

3.1.1 模拟用的高银纬天区类星体和星系样本

通过以下步骤构建的模拟用类星体样本称为 GoodQSO，模拟用星系样本称为 GoodGal。

- 数据源表的选取：

高银纬类星体样本来源于 SDSS_DR16Q_propclean（见 2.5.1; Wu&Shen 2022）和 Gaia DR3 的类星体候选体源表（Gaia_DR3_QSOC; Fu et al. to be submitted）。由于 Gaia DR3 和 SDSS DR16 的天区互补，故可以作为 SDSS_DR16Q_propclean 的补充，以扩大模拟输入的类星体的样本数量。高银纬星系样本来源于 SDSS_DR17_Gal（见 2.5.2）。

- 通过交叉获取测光数据：

  将这两部分数据集整合后，分别根据 ra 和 dec 以 1”的交叉半径和 PS1 DR2、UKIDSS LAS DR9、CatWISE2020 进行交叉，获得 i, z, y, J, H, K, W1, W2 波段的星等。再和 Gaia DR3 交叉获得天体测量数据。

- 对数据集进行校正：

  1) 选取有用的字段并按统一格式命名。
  2) 根据表 1 的优质源筛选判据，筛选出光学-红外波段测光数据优质的源。
  3) 对 J,H,K 和 W1,W2 波段的星等进行校正。由于 PS1 给出的星等是 AB 星等而 UKIDSS 和 Catwise 给出的星等是 Vega 星等，需要将 Vega 星等转换为 AB 星等。按照以下公式[①]：

$$m_{W1[AB]} = m_{W1[Vega]} + 2.699$$
$$m_{W2[AB]} = m_{W2[Vega]} + 3.339$$
$$m_{J[AB]} = m_{J[Vega]} + 0.938$$
$$m_{H[AB]} = m_{H[Vega]} + 1.379$$
$$m_{K[AB]} = m_{K[Vega]} + 1.900$$

[①] 来源：https://wise2.ipac.caltech.edu/docs/release/allsky/expsup/sec4_4h.html#conv2flux

校正后获得数据集 GoodQSO 和 GoodGal，用于后续模拟产生 GoodMockGPQ 和 GoodMockGal。GoodQSO 和 GoodGal 的大小分别为 175175 和 156807。

3.1.2 模拟样本的基本原理和流程

基于宇宙学大尺度上的同质性和各向同性，我们认为通过加减消光模拟得到样本的方法是可行的。这里我们称模拟出来分布在银道面的类星体模拟样本为 GoodMockGPQ, 星系模拟样本为 GoodMockGal。具体流程如下：

- 去除 GoodQSO（或 GoodGal）在原位置处的消光。受星际尘埃的影响，天体的辐射在经过尘埃时受到散射而衰减，对于波段λ,衰减量称为星际消光A_λ，而由星际消光导致的天体星等的变化可以表示为$-\Delta m_\lambda = A_\lambda$。由$R_\lambda = \frac{A_\lambda}{E(B-V)}$得，星等的变化可以通过色余$E(B-V)$和红化率$R_\lambda = \frac{A_\lambda}{A_V} * R_V$计算得到。其中，光学-中红外波段的 $\frac{A_\lambda}{A_V}$ 和 R_V 由 Wang & Chen 2019 给出，R_V=3.1；天区中不同位置天体的$E(B-V)$可以从 Planck 2016 的二维尘埃图（Planck Collaboration et al. 2016）获得，我们通过调用 Python dustmap 库中的 PlanckGNILCQuery 方法实现(Green 2018)。
- 给 GoodQSO（或 GoodGal）分配新的天区位置。随机生成在正负五度天区内均匀分布的位置坐标并随机分配给现有的样本。
- 添加 GoodQSO（或 GoodGal）在新位置处的消光。在新位置处同样根据 Planck 2016 的二维尘埃图计算新位置处的消光，修改星等，得到模拟样本。
- 筛选出位于极限星等区间内的优质模拟样本，得到 GoodMockGPQ（或 GoodMockGal）。为了使得最终得到的模拟样本可以被 PS1、UKIDSS 探测到，我们对模拟样本在新位置处的星等极限进行筛选，使得模拟样本落在 PS1 探测 izy 波段的 5σ区间均值极限星等的范围内（见 2.1）和 UKIDSS LAS 探测 J 波段的 5σ区间极限星等范围内[①]（见 2.2）。相应的星等限制为：

$$imag \leq 23.1$$

$$zmag \leq 22.3$$

$$ymag \leq 21.3$$

$$Jmag_{[vega]} \leq 19.9$$

[①] 注：J 波段的 5σ区间指的是以 2 角秒孔径探测时点源的 Vega 星等的落在探测的 5σ区间内。

这个过程中我们并未对 H, K 波段和 WISE 波段的星等进行限制，原因是：对于 WISE 波段，测光探测的 5σ区间与天区位置有关；对于 H, K 波段，由于红外波段的探测选源相比于光学波段的影响更小，为扩充模拟样本的数量故放宽对 H, K 波段的限制。

- 最后再筛选出在 PS1 探测范围内的样本，即限制$dec > -30°$。

3.1.3 模拟方法的改进——扩充样本容量

以下（3.1.3&3.1.4）以类星体为例进行分析。理论上，由于银心部分消光很强，银心处的类星体被光学-红外波段探测到的概率很小，可以预期经筛选后模拟类星体样本在银心处的分布比较稀疏。由于模拟分配的位置在银道面均匀分布，每一轮中总会有部分被分配到银心位置的 GoodQSO 样本无法通过极限星等的筛选，所以我们无法使得全部的 GoodQSO 样本投入模拟，只能尽可能使得更多的 GoodQSO 样本投入模拟。

设置随机数种子为 1000，对 GoodQSO 模拟一轮后，约有 55%的 GoodQSO 投入模拟。由于模拟一轮得到的 GoodMockGPQ 数量太少，我们考虑通过多轮模拟来扩充样本。

具体方法有两种：

- 无放回抽样模拟：

每模拟一轮后，将本轮未被筛选进模拟样本的 GoodQSO 投入下一轮模拟，以期待在下一轮模拟中重新分配的位置和消光可以使得更多的 GoodQSO 被筛选进模拟样本。注意，这里的无放回抽样并不是统计学意义上的随机抽样，因为我们将样本根据条件筛选这个过程本身并不是随机的，此处方法的命名只是借鉴了无放回的字面概念。

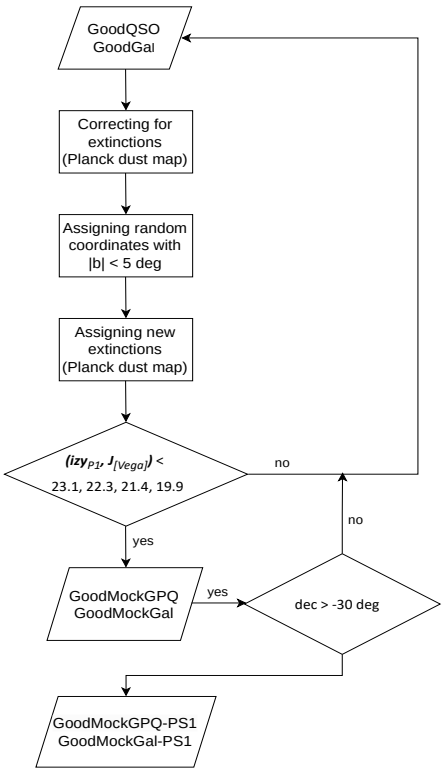

图 3.1 无放回抽样模拟流程图

由于 GoodQSO 中较暗的源更难被筛选进模拟样本，，所以我们预期每一轮中剩余的 GoodQSO 样本的越来越偏暗，所以每一轮被筛选进模拟样本的 GoodQSO 占本轮投入模拟的 GoodQSO 的比例（即每一轮 GoodQSO 的利用率）逐轮递减，GoodQSO 的总利用率涨幅逐轮降低。

通过实际模拟的过程可以看到，每一轮中未被筛选进模拟样本的 GoodQSO 的星等中值逐轮上升，表示越暗的源越难被筛选进模拟样本；每一轮 GoodQSO 的利用率基本上呈逐轮递减。后面的轮数中曲线产生震荡，主要是由于剩余样本较少且质量较差。

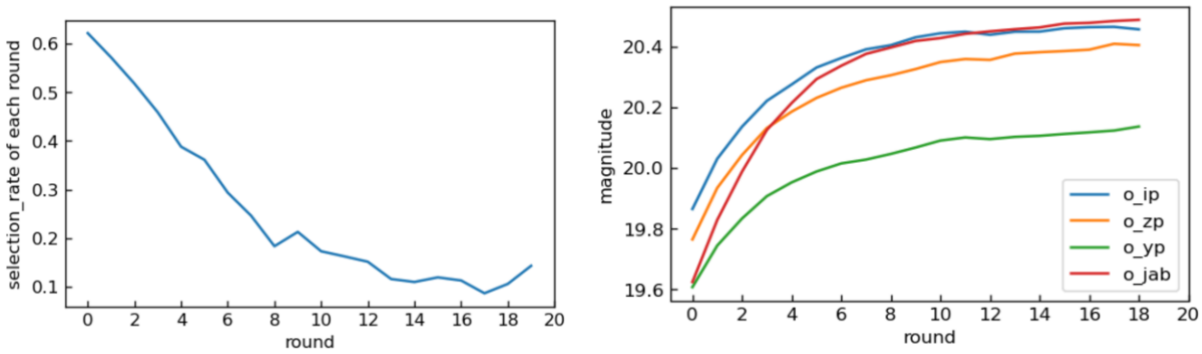

图 3.2 无放回抽样模拟中每轮的利用率和每轮中未被筛选进模拟样本的星等中值

设模拟轮数为 i（第一轮模拟对应 i=0），每轮模拟的随机数种子设为 seed(i+1000)。模拟 10 轮得到的 GoodMockGPQ 的总数为 161021，GoodQSO 利用率为 91.92%

- 有放回抽样模拟(Bootstrap Mocking)

有放回抽样和统计学上的有放回随机采样的概念相同。每一轮模拟中，将本轮被筛选出来的模拟样本和之前筛选出来的所有模拟样本进行 source_id 去重，防止同一个 GoodQSO 样本被重复筛选进 GoodMockGPQ。

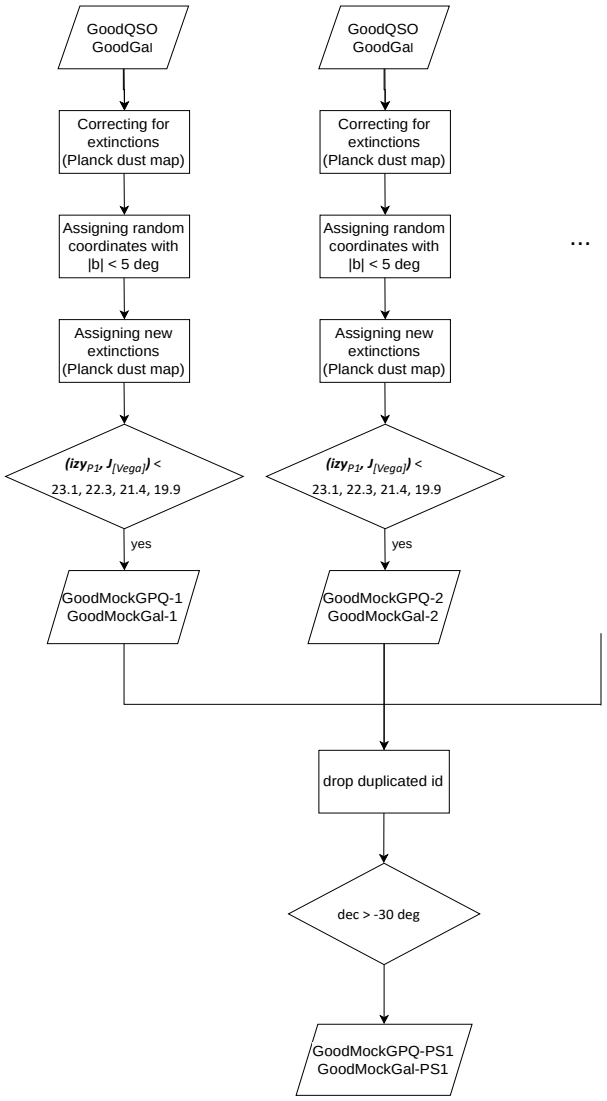

图 3.3 有放回抽样模拟流程图

从理论和实践中都能看出，每一轮模拟的利用率只与随机数种子有关。为方便对照，保持和无放回抽样模拟的随机性一致：设模拟轮数为 i（第一轮模拟对应 i=0），每轮模拟的随机数种子设为 seed(i+1000)。模拟 10 轮得到的 GoodMockGPQ 的总数为 166535，GoodQSO 利用率为 95.07%。

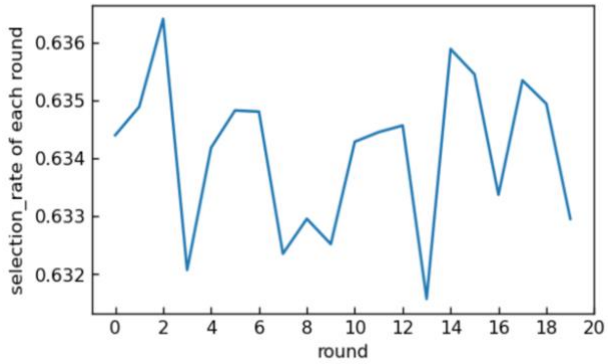

图 3.4 有放回抽样模拟中每轮的利用率

3.1.4 两种模拟方法的比较

从理论上比较，有放回抽样模拟的理论更加明确，随机性比无放回抽样模拟更强，且有放回抽样不存在数据集质量逐轮降低的问题，受数据集的质量分布影响较小。

从实际模拟过程上比较，同样的随机数种子和同样的模拟轮数，显然有放回抽样模拟得到的 GoodMockGPQ 更多，所以更加省时省力。

3.1.5 模拟结果

对 GoodQSO 和 GoodGal 均用有放回抽样进行模拟。可以看出，由于我们输入模拟的星系样本整体比类星体亮很多，所以达到同样利用率星系所需的模拟轮数更少。

表 3.1 GoodQSO 和 GoodGal 有放回抽样进行模拟的结果

	输入样本的数量	GoodMock Samples 数量	模拟利用率
GoodQSO	175175	166535	模拟十轮：95.07%
GoodGal	156805	156794	模拟十轮：99.99%

筛选出在 PS1 探测范围内（$\mathrm{dec} > -30°$）的样本后，GoodMockGPQ 和 GoodMockGal 的 Healpixmap 天区分布如下图所示[①]

[①] 这里的 Healpixmap 使用 Nside=64。

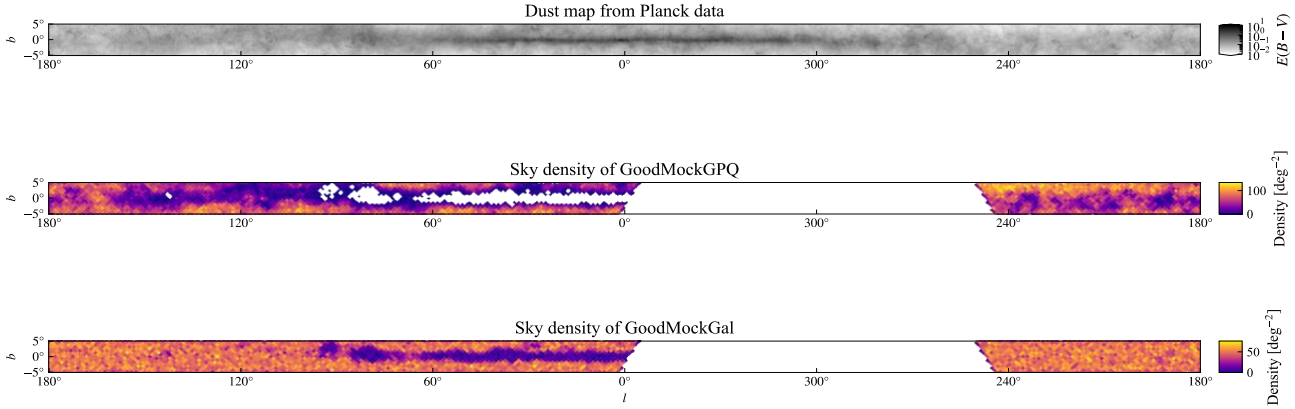

图 3.5 GoodMockGPQ 和 GoodMockGal 的天区密度分布与 Planck Dustmap 的对照

在模拟的过程中，随机分配坐标使得模拟的源均匀分布在指定天区内，而根据 Planck Dustmap 对照可得，靠近银心处的消光更加严重，所以在经过星等极限的筛选后，银心处的模拟样本显著减少。

3.2 类星体先验概率的计算

为了后续计算类星体的后验概率时减小机器学习的偏差，我们需要对实际天区中类星体的密度分布，即类星体的先验概率分布进行估算。而实际天区中类星体的分布和 GoodMockGPQ 的天区分布并不一致，因为我们模拟样本的过程中并没有计入银道面天区中天体的拥挤程度对测光质量的影响，且 GoodMockGPQ 的天区密度很大程度上取决于我们模拟的方法和轮数。因此，若想得到实际天区中类星体的先验分布，我们需要对 GoodMockGPQ 的分布做两方面的校正：

设正负五度天区中所有的 PS1 X UKIDSS GPS X CatWISE（以下简称 GPC）的源密度分布为D_{all}，根据表 1 中的判据筛选出来的优质源的密度分布为D_{goodph}，故我们可以用$\frac{D_{goodph}}{D_{all}}$来衡量天区中天体的拥挤程度对测光质量的影响。设 GoodMockGPQ 的密度分布为D_{old}，则$D_{new}{}' = D_{old} * \frac{D_{goodph}}{D_{all}}$可以用来表示考虑到天体的拥挤程度时 GoodMockGPQ 的相对密度分布。

为了得到类星体的绝对密度分布，我们还需要剔除模拟的具体过程对 GoodMockGPQ 的影响。设 GoodQSO 的密度分布为$D_{GoodQSO}$，我们可以用$\frac{Median(D_{GoodQSO})}{Median(D_{new}{}')}$来校正 GoodMockGPQ 和实际天区中类星体分布的偏移。注意，由于 GoodQSO 的样本并不完备，

所以 Median($D_{GoodQSO}$)比实际上高银纬天区类星体密度的中值小很多，因此用 $\frac{Median(D_{GoodQSO})}{Median(D_{new}')}$校正后得到的类星体的密度分布实际上只是一个下限，即正负五度天区内类星体的实际密度分布≥ $D_{new}' * \frac{Median(D_{GoodQSO})}{Median(D_{new}')}$。我们设这个下限为$D_{new}$。

我们将全天按照 Nside=256 分成 786432 个 Healpix 像素，根据天体的 ra,dec 坐标将每个天体分到对应的像素中，从而统计出每个像素中天体的密度（即位于该像素中天体的总数除以该像素对应的天区的面积），形成天体密度按像素的分布 D。

经计算，D_{new}的中值为 56.76 $\deg^{-2}$, 最大值为 253.17 $\deg^{-2}$[①]。各密度分布计算情况如下图所示：

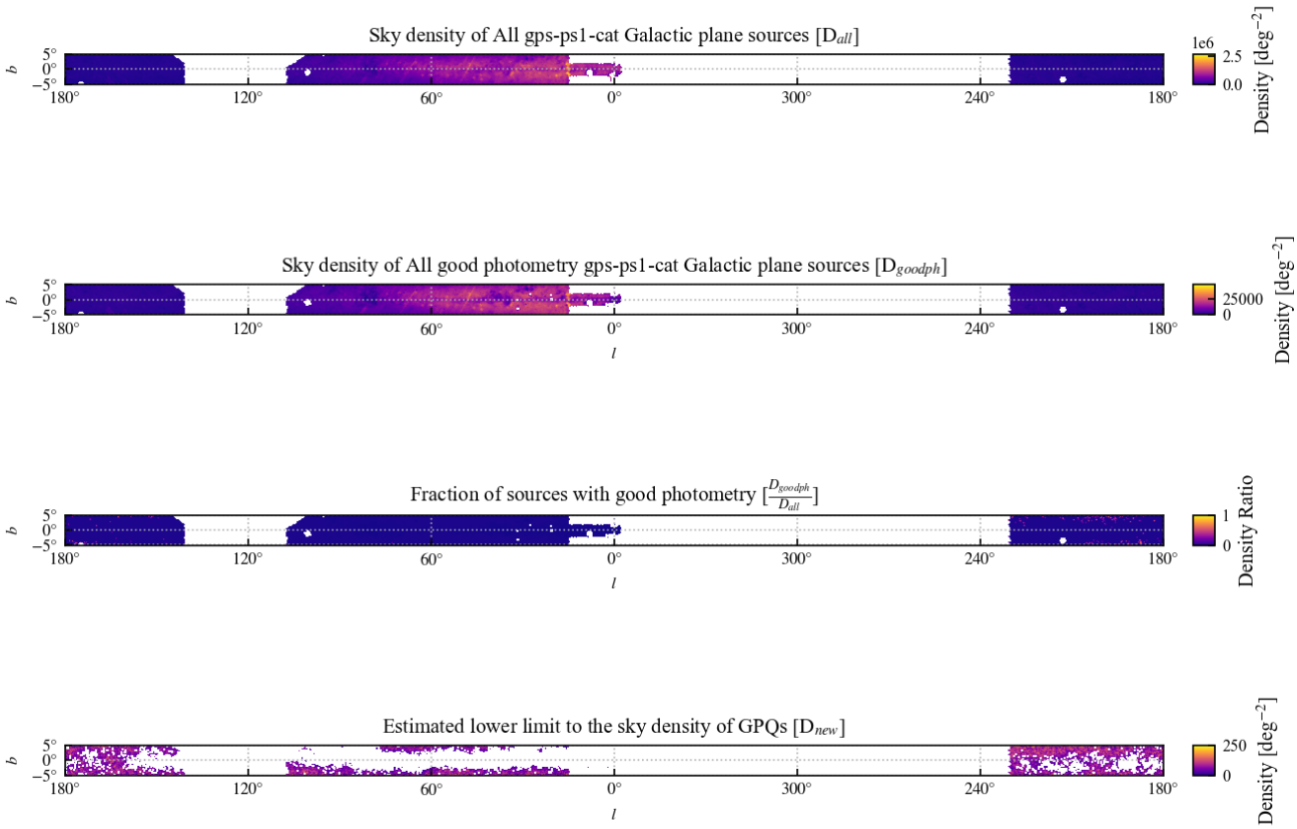

图 3.6 计算类星体先验概率分布所用的先验天区密度分布

根据正负五度天区中类星体和 GPC 源的密度分布，我们可以计算出正负五度天区内类星体的概率分布为$P_{prior} = \frac{D_{new}}{D_{goodph}}$，即我们对类星体先验概率的估计。$P_{prior}$的分布如下：

① 注：我们取没有天体的像素的密度为缺失值 np.nan，这里中值和最大值的统计均不计入缺失值的像素。

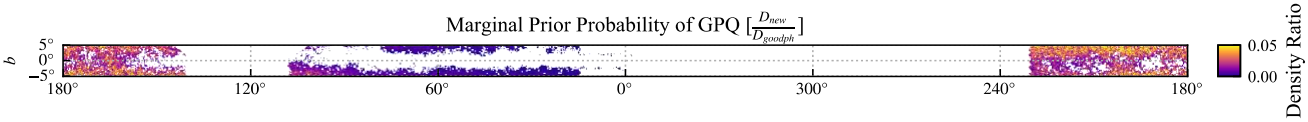

图 3.7 类星体先验概率分布

值得一提的是，为了尽可能地平滑天体密度的分布，以最大限度地减轻 Healpixmap 边界天体密度异常大的问题，我们将 Nside 从 64 调高到 256 以切割出更多的全天像素，使得天区密度分布的连续性和精确性更强。然后我们通过查看P_{prior}在的天区分布，发现P_{prior}异常大的像素基本都分布于 Healpixmap 的边界处，于是我们手动将位于 99%分位区间以外（即$P_{prior} > 0.0628\ \mathrm{deg}^{-2}$的像素）的$P_{prior}$设置为正负五度天区的中值$0.0130\ \mathrm{deg}^{-2}$, 以最大限度地减轻边界异常值对后续类星体后验概率计算的影响。

3.3 恒星样本的构建

我们从 Gaia DR3 中选取 OBA 型星和 FGKM 型星、以及从不同巡天中选取矮星和亚矮星来构建恒星的训练样本。

为了获得天体物理学参数质量更高、更纯净、更有代表性的源，我们最初计划从 Gaia DR3 的 Golden Sample 子集（Creevey et al. 2022）里分别获取 OBA 型星和 FGKM 型星的源表。之后我们发现，Gaia DR3 在选取 Golden Sample 子集时对 FGKM 型星的限制比 OBA 型星严格的多，导致 Golden Sample 里的 FGKM 型星远少于 OBA 型星，然而这与实际天区中恒星的数量分布完全不同，所以这样选源构成的样本会造成数据集的偏移。因此我们选择重新从 Gaia DR3 的总表里选取 FGKM 型星，并加严 Golden Sample 里 OBA 型星的筛选判据，以使得我们选取的恒星样本的赫罗图分布和实际天区更加接近。我们通过 ADQL Query 从 Gaia Archive 网站选取 OBA 和 FGKM 型星样本及所需的参数。

- OBA 型星：我们直接选取 Golden Sample 子集里的 OBA 型星，按 source_id 回交叉到 Gaia DR3 的总表，并和 Gaia DR3 的 astrophysical_parameters 字段交叉获得天体物理学参数、和 ps1 交叉获得光学波段的星等，再通过更加严格的限制进行进一步筛选，以提高 OBA 型星样本的纯净度。具体的限制条件如下表：

表 3.2 OBA 型星样本筛选判据

判据	物理意义
vtan_flag = 0	切向速度小于 180km/s，可剔除 halo stars 的污染
ruwe < 1.4 & ipd_frac_multi_peak < 6 & classprob_dsc_combmod_star > 0.9	可以剔除双星的污染
teff_gspphot > 7000	限制有效温度
phot_bp_n_obs > 19 & phot_rp_n_obs > 19 & phot_g_n_obs > 150	对 BP/RP/G 波段测光有贡献的观测的次数
parallax_over_error > 15	限制视差的信噪比大于 15
phot_bp_n_blended_transits < 10	限制 BP 波段为多个源叠加的 transits 数量
(lum_flame_upper- lum_flame_lower)/lum_flame < 0.2	限制 lum_flame 的误差在 0.2 以内
b>-5 & b<5	限制在银纬正负五度天区

• FGKM 型星： Gaia DR3 在构建 FGKM 型星的 Golden Sample 时应用的判据为 FGKM_3，相比于判据 FGKM_2 增加了许多额外的筛选，使得 Golden Sample 中的 FGKM 型星数量过少。这里我们基本根据判据 FGKM_2 重新选取 FGKM 型星的样本，具体限制条件如下：

表 3.3 FGKM 型星样本筛选判据

判据	物理意义
ruwe < 1.4 & ipd_frac_multi_peak < 6 & classprob_dsc_combmod_star > 0.9	可以剔除双星的污染
2500<teff_gspphot <7500 & teff_gspphot_upper- teff_gspphot_lower < 150	限制有效温度

mh_gspphot > -0.8	限制$\log_{10}\frac{\text{铁丰度}}{\text{氢丰度}} > -0.8$
mg_gspphot < 12	限制 G 波段星等<12
logposterior_gspphot > -4000	限制和统计模型的契合度
radius_gspphot < 100	恒星半径小于 100
libname_gspphot='MARCS' | libname_gspphot='PHOENIX'	只有'MARCS'或'PHOENIX'这两个恒星 libraries 中选取的恒星才适用于这里的限制条件
phot_bp_n_obs > 19 & phot_rp_n_obs > 19 & phot_g_n_obs > 150	对 BP/RP/G 波段测光有贡献的观测的次数
parallax_over_error > 15	限制视差的信噪比大于 15
phot_bp_n_blended_transits < 10	限制 BP 波段为多个源叠加的 transits 数量
(lum_flame_upper-lum_flame_lower)/lum_flame < 0.2	限制 lum_flame 的误差在 0.2 以内
b>-5 & b<5	限制在银纬正负五度天区

以上方法选取出来的 OBA 型星和 FGKM 型星在赫罗图上的分布如下图所示。

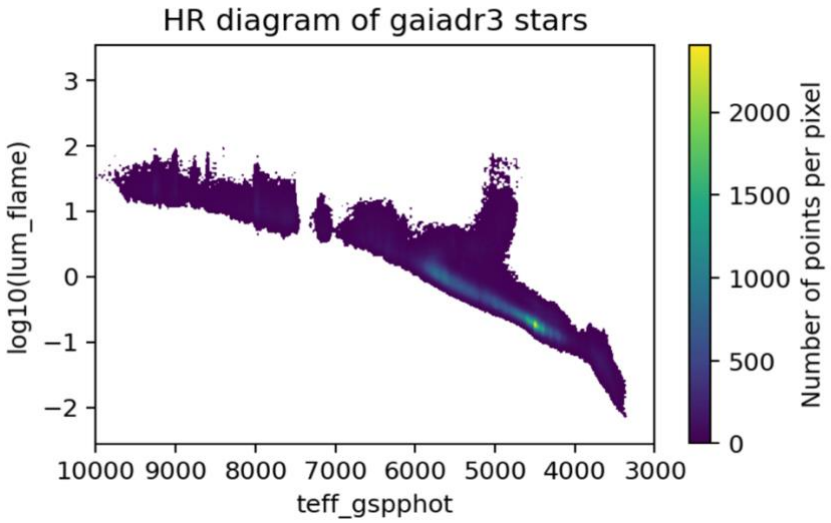

图 3.8 OBA 型星和 FGKM 型星样本在赫罗图上的分布

其中 teff_gspphot<7500K 的部分是 FGKM 型星，teff_gspphot>7000K 的部分是 OBA 型星。可以看出赫罗图的分布和实际天区更加相符。

矮星和亚矮星样本方面，我们选取了主要在光学波段可见的 SDSS DR7 的 M 型矮星（West et al. 2011）、Gaia DR3 的超低温矮星（见 2.4）、LAMOST 的 M 型矮星（Li et al. 2021）、LAMOST 的超低温矮星（Wang et al. 2022）、SVO 的晚型亚矮星（Lodieu et al. 2017），以及主要在红外波段可见的矮星（Kirkpatrick et al. 2019）和亚矮星（Zhang et al. 2018）。将这些矮星和亚矮星样本整合到一起和 PS1 DR2, UKIDSS LAS, CatWISE2020 交叉后得到各波段的测光数据。注意，由于这些矮星和亚矮星样本基本都分布在高银纬，所以我们选择和 UKIDSS LAS 交叉而不是和 GPS 交叉。

然后我们通过两步筛选出测光优质的样本集：根据表 1 选出测光质量高的源，再根据 2.1，2.2，2.3 选出同时在 PS1 i,z,y, UKIDSS J,H,K[①], CatWISE2020 W1,W2 的极限星内的源。最后我们剔除不在 PS1 探测范围内（$dec < -30°$）和在 Gaia DR3 类星体或星系候选体中的源，以获得更加纯净的恒星样本。

3.4 训练样本的特征分析

将 3.1 和 3.3 构建的类星体、星系和恒星样本进行整合，标注上每一个样本所属的分类标签，构成正负五度天区的训练样本。训练样本中恒星、类星体和星系的比例为 476373:120891:111681。

以下我们画出训练样本在颜色空间的分布：

[①] 注：OBA 和 FGKM 型星使用 UKIDSS GPS 的判据，矮星和亚矮星使用 UKIDSS LAS 的判据

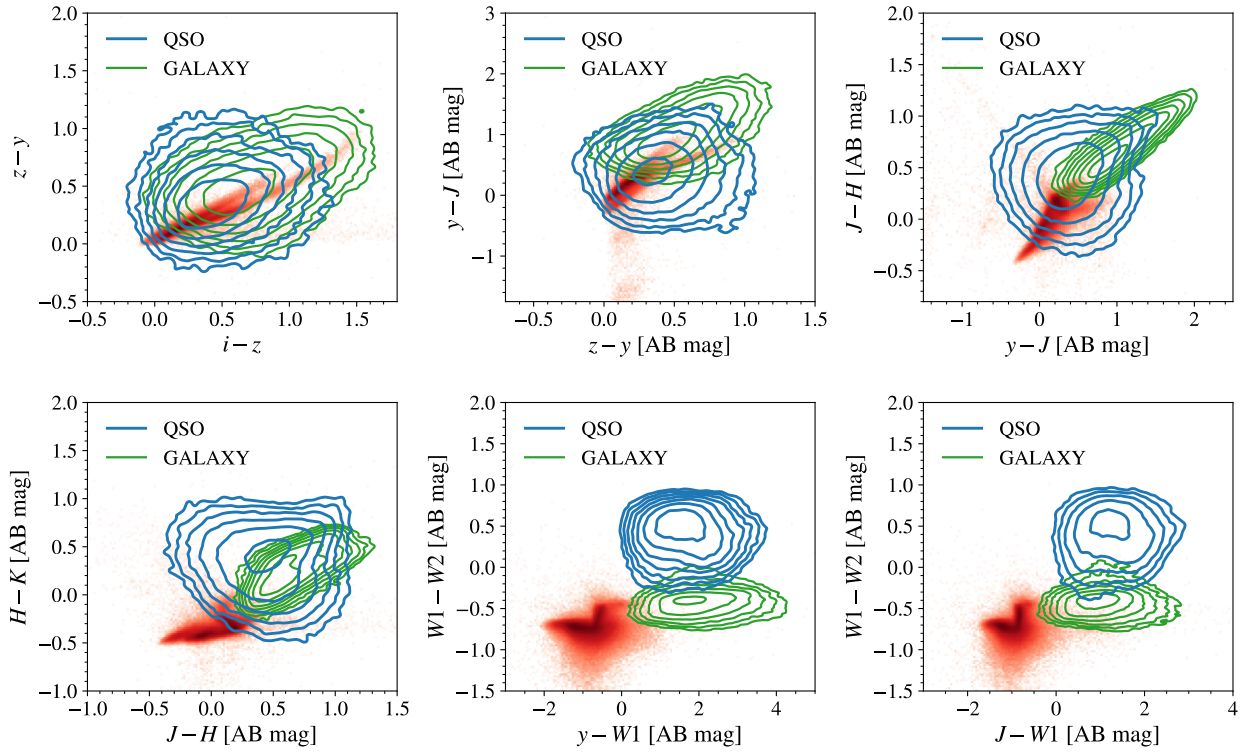

图 3.9 训练样本在颜色空间的分布。

我们分别画出光学和近红外波段依次相减得到的颜色图。对于光学/近红外与 WISE 中红外波段相减的图，我们分别选择探测最好且三种类别天体分布区别最明显的光学(y)和近红外(J)波段构成的颜色图呈现。上述颜色分布也会运用到后续测试集 Pred_GPQ 的筛选中（见 5.2）

从图中可以看出，y-W1～W1-W2 和 J-W1～W1-W2 的颜色分布图上，三类天体的区别最为明显，这也表明了 W1-W2，y-W1，J-W1 这三个特征对于区分这三类天体的重要性（与后续机器学习分类模型评估出的特征重要性也相吻合，见 4.4）。JHK 波段的颜色图中类星体与恒星样本的重合稍多，但我们仍然可以据此筛选掉左下方大量恒星样本。

第四章 机器学习分类模型

4.1 XGBoost 的介绍

我们选用机器学习中的 XGBoost(Chen & Guestrin 2016)来执行银道面类星体选源的分类任务。XGBoost 全名 eXtreme Gradient Boosting，是一种梯度提升算法，用于解决有监督学习问题。XGBoost 的模型中，每一步的决策树可以纠正上一棵决策树的损失，从而达到逐步优化目标函数的效果。相比于传统的梯度提升模型（GBM），XGBoost 在许多方面做了优化：

- XGBoost 在目标函数表达式中加入正则项，可以有效地控制模型复杂度，减轻模型的过拟合问题，使得模型泛化能力更强；
- XGBoost 可以处理缺失值问题，对稀疏数据集更加友好。

XGBoost 在天文学中的应用日益广泛，例如根据光谱巡天分辨 M 型巨星和 M 型矮星（Yi et al. 2019）、从 3FGL 源表（Acero et al. 2015）中选取星系候选体（Mirabal et al. 2016），其解决天文相关问题的优良的性能也得到越来越多的体现。

4.2 分类特征的选取和原因

我们一共选取 12 个测光颜色作为机器学习分类的特征：i-z, z-y, y-J, J-H, H-K, i-W1, z-W1, y-W1, J-W1, H-W1, K-W1, W1-W2。其中，光学波段的 i-z, z-y 可以减少颜色图上类星体在恒星位点处的重叠；近红外波段的颜色 y-J, J-H, H-K 可以用来更好地区分出高红移类星体，因为高红移类星体的光学波段消光严重；W1-W2 可以用来更好地从星系和恒星中筛选类星体，因为类星体的 W1-W2 波段颜色比星系和恒星要红；Jin et al.(2019)的研究表明 i-W1,y-W1,z-W2 这三个颜色对 XGBoost 区分类星体和恒星有很大的帮助，由于 W1 比 W2 波段的测光更加精确，我们据此构建 i-W1, z-W1, y-W1, J-W1, H-W1, K-W1 六个颜色特征，以比相邻波段颜色相减更宽的波长范围来描述天体的光谱能量分布。我们并未使用 W3 波段的数据，因为 W3 波段的数据质量较差，信噪比较高且缺失值也较多。

4.3 分类流程

为了减轻数据集类别不平衡的偏移，我们将三分类问题拆解成两步二分类问题，第一步从训练样本中分类出恒星和河外天体，相应的分类模型设为 clf-1，第二步从河外天体中

分类出类星体和星系，分类模型设为 clf-2。因此，我们将总样本拆分成河外天体与恒星、类星体和星系两个集合，分别作为两步分类的训练样本。

我们采取 5 折交叉验证，即将训练样本分割成 5 个子集，其中一个子集用做验证集，剩下 4 个子集用来训练，这样 5 个子集分别轮流做一次验证集，最后平均 5 次训练的评估结果，以减少如过拟合和选择偏差的问题。同时，为了优化分类模型的超参数，我们采用 Optuna (Akiba et al. 2019)进行调参，使得 500 次采样训练里模型的对数损失函数最小。对数损失函数的定义为：设某一个样本的真实标签为$y(y \in \{0,1\})$,该样本被模型预测为类别$y = 1$的概率为p,则该样本的对数损失函数为$\log_{loss}(y,p) = -(y\log p + (1-y)\log(1-p))$。

在调优超参数时，我们固定提升轮数（boosting rounds）num_boost_round=100，因为学习率 eta 表征优化过程向损失函数前进的步长，学习率的增加会使得提升轮数相应减少，故提升轮数和学习率不可同时进行超参数调优。最终得到两步分类的最优超参数如下：

表 4.1 Optuna 调优的两步分类最优超参数

	clf-1	clf-2
eta	0.3	0.1
lambda	2.9077613895719	0.8545938248212175
alpha	1.5263290780459133	0.4512391408998466
max_depth	9	8
gamma	1.0003397864024768	0.2038341012648358
grow_policy	depthwise	depthwise
min_child_weight	4	1
subsample	0.9129317176922225	0.9094012577087329
colsample_bytree	0.9886871929591262	0.8079200202937279
max_delta_step	6	2

在正式训练分类模型的过程中，我们将训练样本按照 4:1 的比例划分为训练集和验证集，其中 4:1 的比例也与调优超参数过程中的 5 折交叉验证相匹配。我们固定分类模型的参数 objective=binary:logistic、booster=gbtree、tree method=hist，然后传入调优的超参数（除

了 eta），并通过减小 eta、增加 n_estimators（即 num_boost_round）来减小分类模型的泛化误差，最后根据模型的评估参数（见 4.4）选出最优的 eta 和 n_estimators，以此构建最终的机器学习分类模型。

4.4 分类模型的评估

在机器学习训练的过程中，我们采用以下六种参数来评估分类模型：正确率(Accuracy)，精确率（Precision），召回率（recall），F1 分数，MCC（Matthews correlation coefficient），和 AUCPR（PR 曲线下的面积）。定义真阳性样本的个数为 TP，假阳性为 FP，真阴性为 TF，假阴性为 FN，则这些参数可以表达为：

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

$$MCC = \frac{TP * TN + FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

$AUCPR$可以通过对 PR 曲线积分获得。

这六个参数中，Accuracy, Precision, Recall 和 F1 是衡量机器学习模型的常用指标。但在类别不平衡的情况下，这四个指标对模型的评估会严重偏向数量更大的类别，所以我们引入 MCC 和 AUCPR 进行补充。MCC 将 TP,TN,FP,FN 四个指标都考虑进去，不受数据集中两种类别占比的影响(Chicco & Jurman 2020)；而在类别不平衡的情形下，PR 曲线比 ROC 曲线更能反映模型的性能(Davis & Goadrich 2006; Saito & Rehmsmeier 2015)，我们可以通过 PR 曲线与坐标轴围成的面积 AUCPR 来衡量模型的 PR 曲线。

在正式训练的过程中，我们通过衡量模型这六个方面的参数，选出使模型评估指标最优的 eta 和 n_estimators， 得到 clf-1 的 eta=0.05, n_estimators=1000, clf-2 的 eta=0.02, n_estimators=1000。将两步分类过程中使用默认超参数的模型和使用调优超参数的模型进行评估对比（其中默认超参数的模型和使用调优超参数的模型都是采用最优的 eta 和 n_estimators.具体如下表：

表 4.2 clf-1 使用默认超参数和使用调优超参数的评估参数对比

评估参数	clf-1 Default	clf-1 Optimized
Accuracy	0.999147	0.999189
Precision	0.999496	0.999538
Recall	0.999233	0.999255
F1	0.999365	0.999396
MCC	0.998065	0.998161
AUCPR	0.999990	0.999991

表 4.3 clf-2 使用默认超参数和使用调优超参数的评估参数对比

评估参数	clf-2 Default	clf-2 Optimized
Accuracy	0.987337	0.987703
Precision	0.987795	0.987844
Recall	0.987836	0.988498
F1	0.987816	0.988171
MCC	0.974636	0.975367
AUCPR	0.998354	0.998521

同时，使用 Optuna 调优过的分类模型得到混淆矩阵分别如下图：

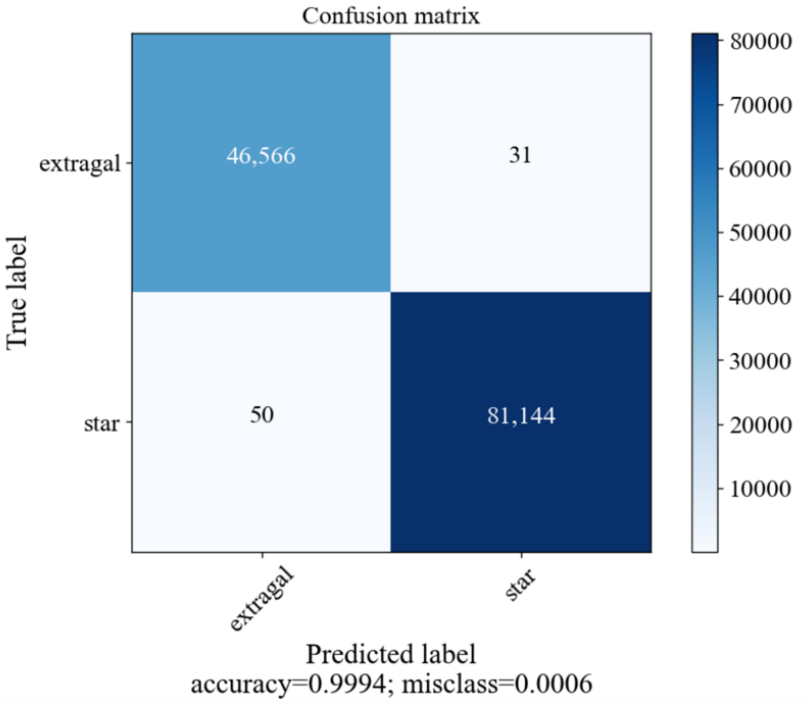

图 4.1 clf-1 的混淆矩阵

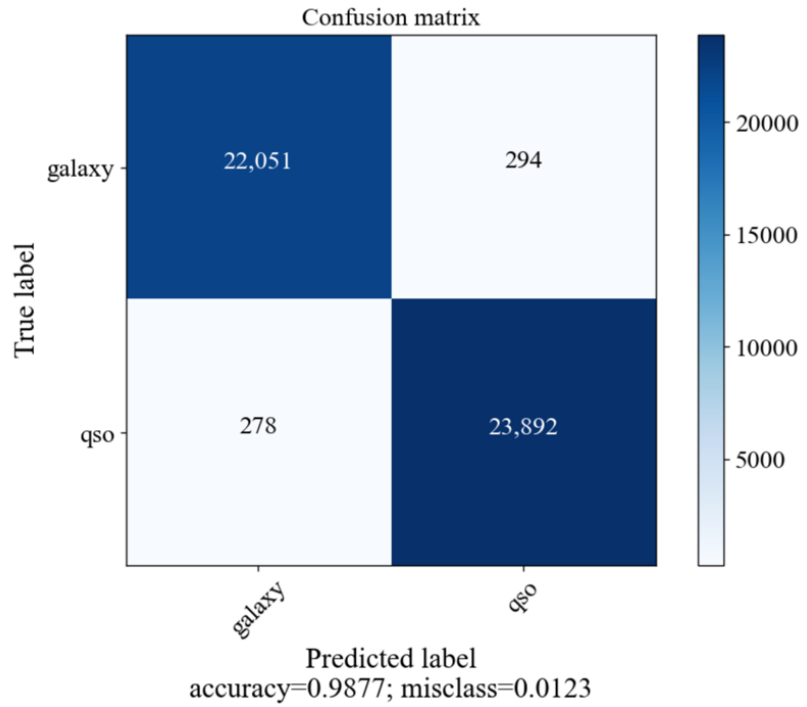

图 4.2 clf-2 的混淆矩阵

可以看出，分类模型在验证集上的表现非常不错，且经过 Optuna 调优过的分类模型在各个评估指标上都表现更好。由于类星体和星系类别特征的重叠更大，第二步分类比第一步分类更难，所以 clf-2 的表现不如 clf-1 好。

训练过程中，还对各个分类特征的重要性做了评估。两个重要的评估指标分别为频率 Frequency 和增益 Gain，其中频率表示该特征在决策树中作为分割样本的节点出现的频率，增益表示在决策树中该特征被用作分割样本的节点时带来的目标函数优化的平均值，二者都是表征特征重要性的重要指标。需要注意的是，增益是更常用的指标，更能代表一个特征的有无对模型效果的影响。由于我们的分类特征都是数值特征，没有枚举特征，所以这里没有讨论特征重要性中的覆盖度 Cover。

两个分类模型的特征重要性如下图：

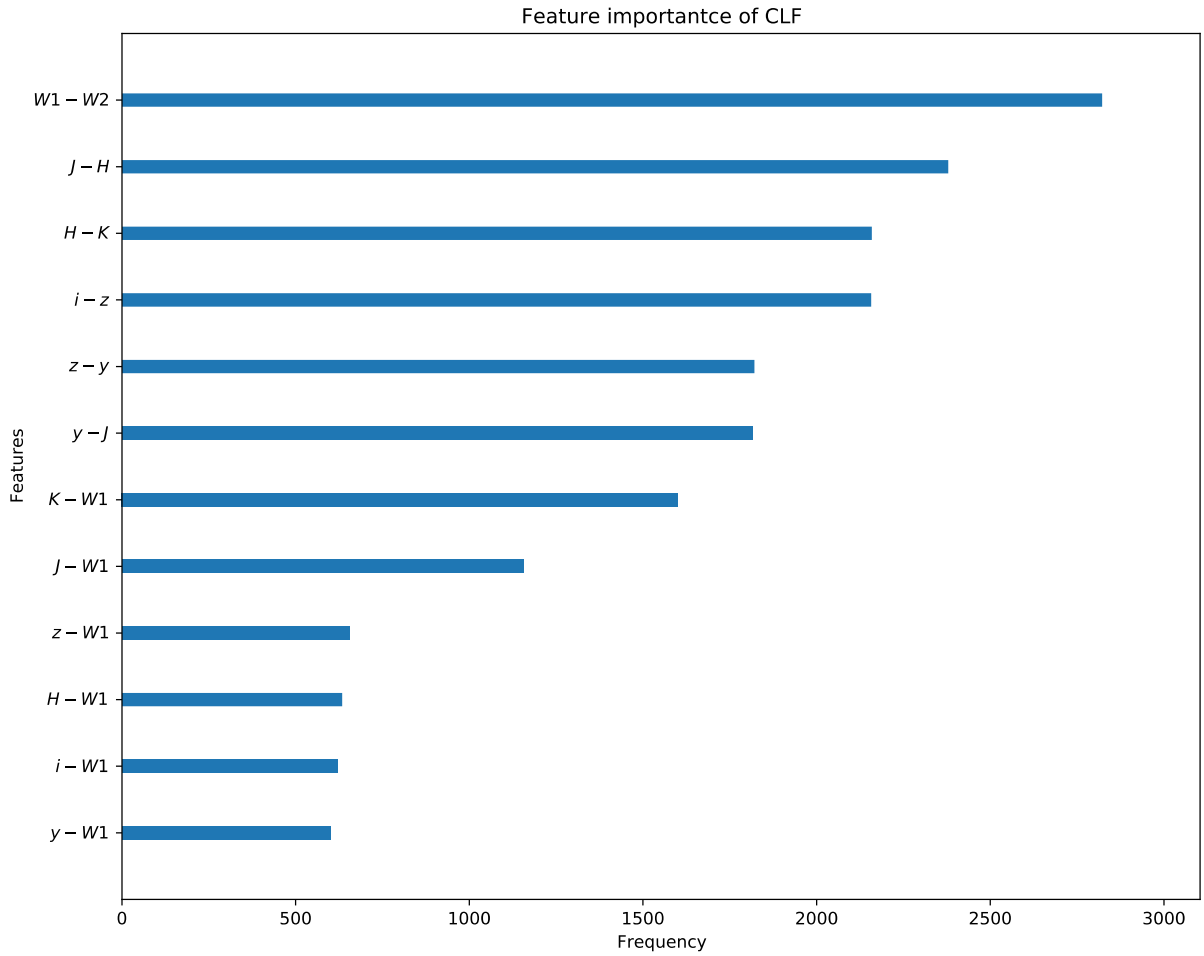

图 4.3 clf-1 的特征重要性-频率

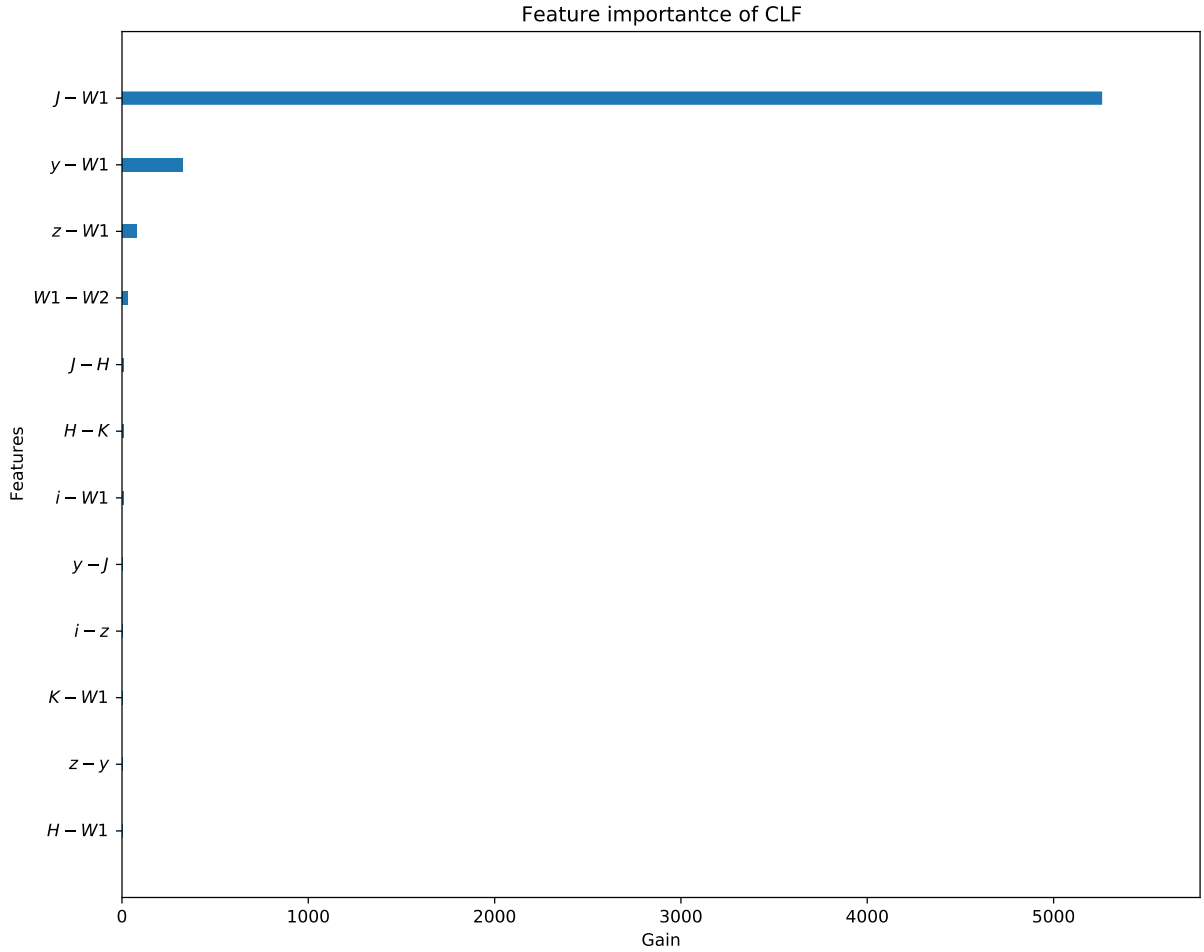

图 4.4 clf-1 的特征重要性-增益

clf-1 中 J-W1 特征的增益断层最大，其次是 y-W1，说明光学-近红外波段的测光数据对于分类恒星和河外天体的影响非常大。

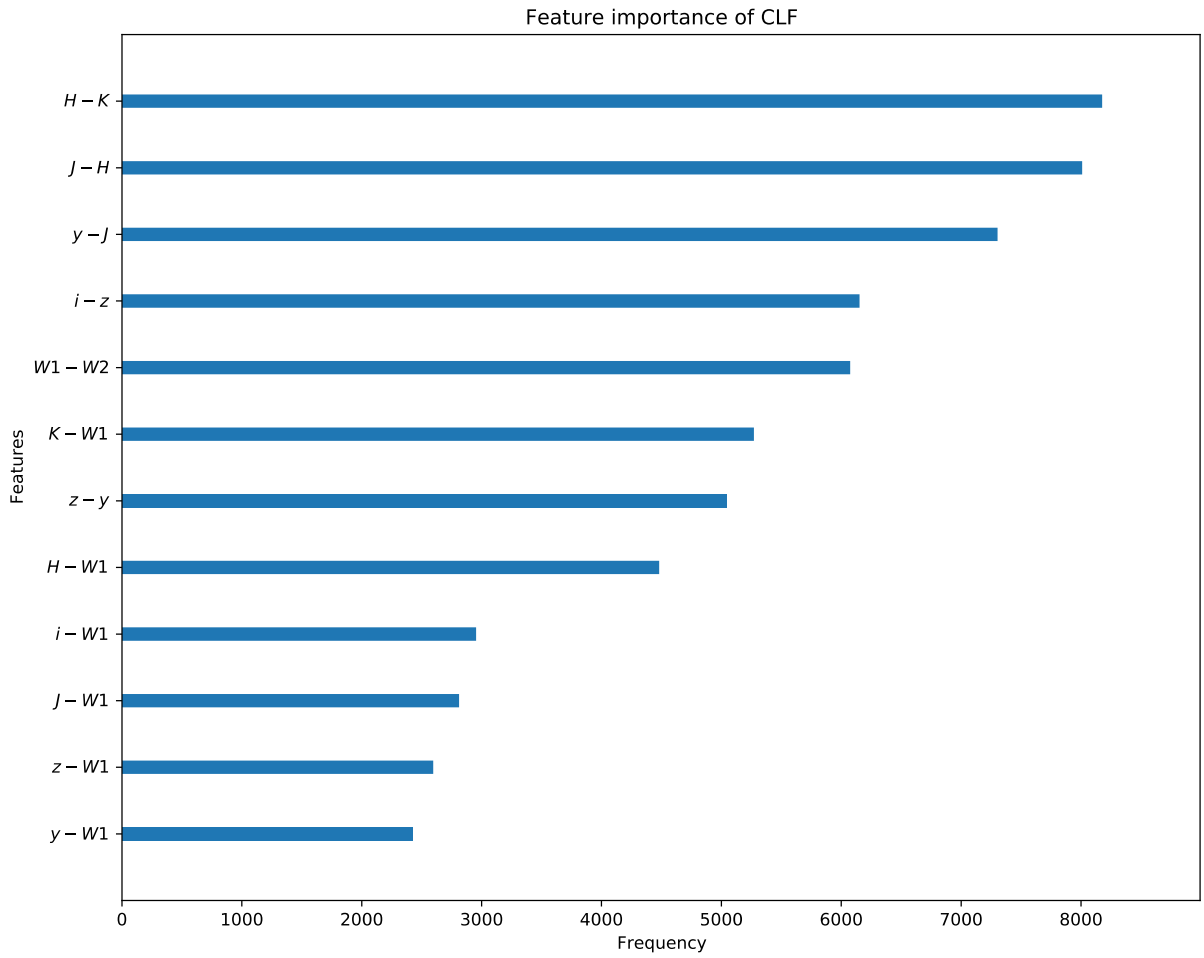

图 4.5 clf-2 的特征重要性-频率

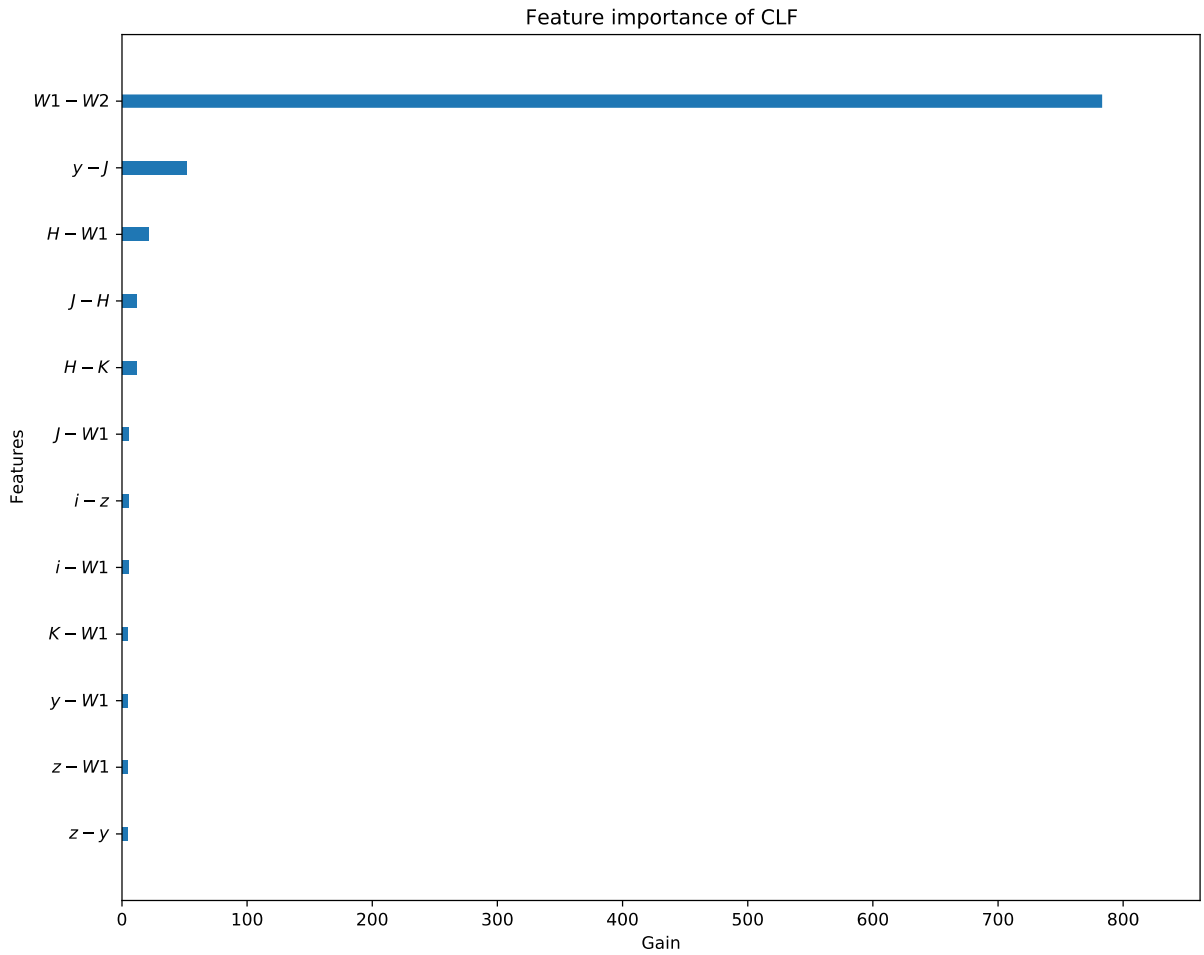

图 4.6 clf-2 的特征重要性-增益

clf-2 中 W1-W2 特征的增益最大，说明 W1-W2 这一特征对区分类星体和星系最为重要。

第五章 类星体候选体的筛选

5.1 分类模型在测试集上的表现

正负五度天区的总测试集由 PS1, UKIDSS GPS, CatWISE2020 交叉获得，即 3.2 中所提到的 GPC 源，记为 All，再通过表 2 的判据和极限星等筛选出优质测试样本。

All 文件中含有10^9数量级的源，为了便于处理，我们将其分割为基本均等的 979 个子文件，分别对每个子文件进行机器学习分类预测，最后再将预测出的类星体即其对应的预测概率整合成一张表以进行后续的计算。

由于正负五度天区恒星的污染非常严重，为了提高分类预测出的类星体候选体的纯净度，我们设置了以下分类阈值：将第一步分类为河外天体的概率 P1 大于 99%的样本筛选出来进行第二步分类，再将第二步分类中分类为类星体的概率 P2 大于 99%的样本筛选出来，构成我们机器学习分类预测出的类星体候选体。经过整合，我们从 All 的 325,673,408 个源中选出 1,145,937 个类星体，构成集合 PredGPQ，其中每一个样本的总预测概率 P=P1*P2。

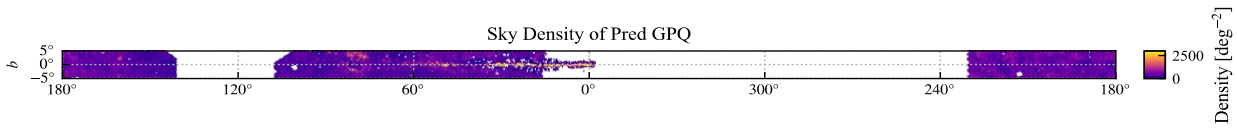

图 5.1 PredGPQ 的天区密度分布

但根据上图显示， PredGPQ 反而在银心处密度很高，说明其中包含大量的恒星污染。我们画出 PredGPQ 在颜色空间的分布，并与 3.1 中的类星体训练样本和 3.3 得到的恒星训练样本比较如下：

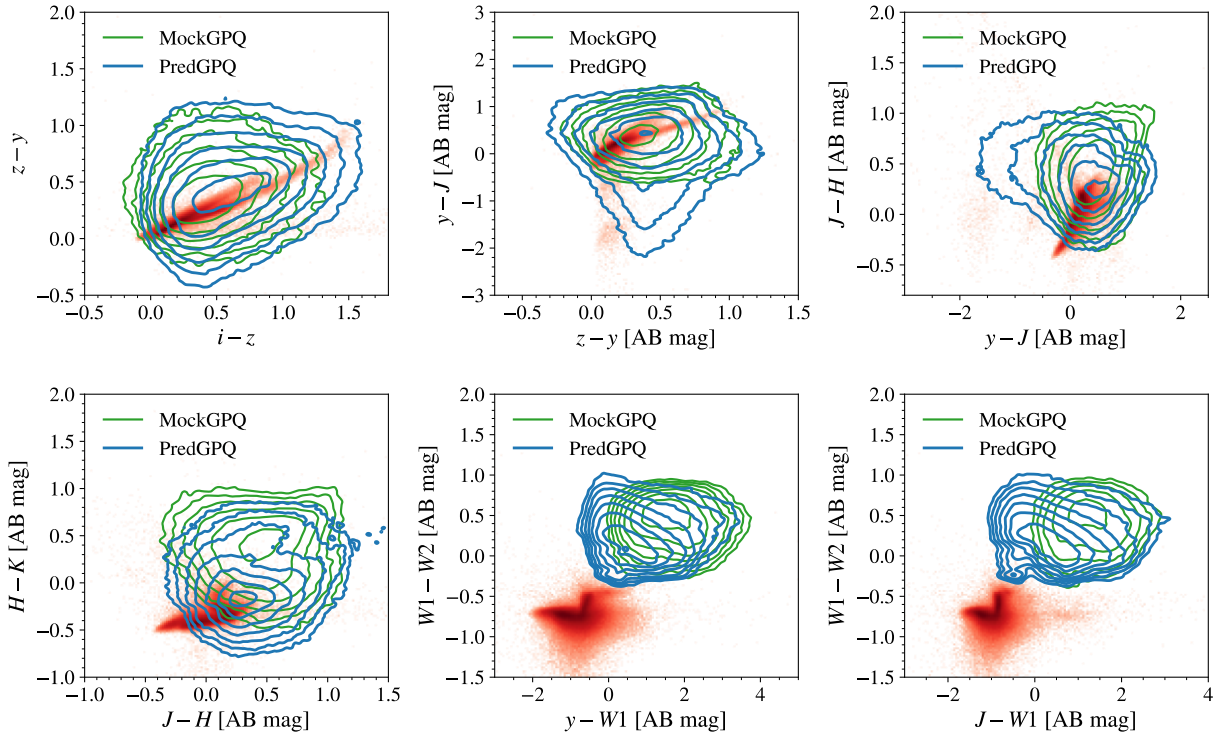

图 5.2 PredGPQ 的颜色空间分布

从下方三张子图可以看出，PredGPQ 的红外波段颜色更靠近恒星位点，说明 PredGPQ 中恒星的污染比较严重；从上方三张子图可以看出，PredGPQ 的光学波段颜色朝恒星样本分布的弥散较大，这同时也说明恒星训练样本的弥散会降低机器学习分类模型的效果；

为了进一步提纯类星体候选体，我们需要进行后验概率的校正和多种截断方法的筛选。

5.2 类星体后验概率的计算

由于机器学习训练分类模型时，我们输入的训练样本的分布和实际天区中天体的分布并不一致，所以机器学习的预测结果是有偏差的(biased)。为此我们需要引入实际天区中天体的分布情况，即先验分布，来对机器学习的预测结果进行校正。

设机器学习的特征集合为 x，类别集合为 y。据此设机器学习预测的概率为$P_{biased}(y|x)$，表示分类模型将每一个已知特征 x 的源预测为类别 y 的概率；设似然(likelihood)为$P(x|y)$，表示已知样本类别为 y 其特征分布为 x 的概率；设训练样本的先验分布为$P_{biased}(y)$，表示类别 y 在训练样本中的占比。设校正过的类星体后验概率为$P_{real}(y|x)$，表示实际天区中每一个已知特征 x 的源为类别 y 的概率；似然(likelihood)仍然为$P(x|y)$不变；设实际天区中天体的先验分布为$P_{real}(y)$，表示类别 y 在实际天区中的先验分布。根据贝叶斯公式：

$$P_{biased}(y|x) \propto P(x|y) * P_{biased}(y)$$

$$P_{real}(y|x) \propto P(x|y) * P_{real}(y)$$

$$\text{得，} P_{real}(y|x) \propto \frac{P_{biased}(y|x) * P_{real}(y)}{P_{biased}(y)}$$

其中，由 3.4 得$P_{biased}(y) = \frac{120891}{638951} \cong 0.17$ ；$P_{real}(y)$即 3.2 中类星体的先验概率分布P_{prior}，与 Healpixmap 中的每个像素一一对应；按照 5.1 的模型分类阈值，$P_{biased}(y|x)$即 Pred_GPQ 的总预测概率 P，与每个样本一一对应。我们将 Pred_GPQ 中每一个样本按照其 ra,dec 坐标计算出其在 Healpixmap 中所处的像素，得到每一个样本所在像素的$P_{real}(y) = P_{prior}$，然后对每一个样本计算$\frac{P_{biased}(y|x) * P_{real}(y)}{P_{biased}(y)} = P * P_{prior}/0.17$，便可以得到每一个样本为类星体的后验概率$P_{real}(y|x)$的分布。

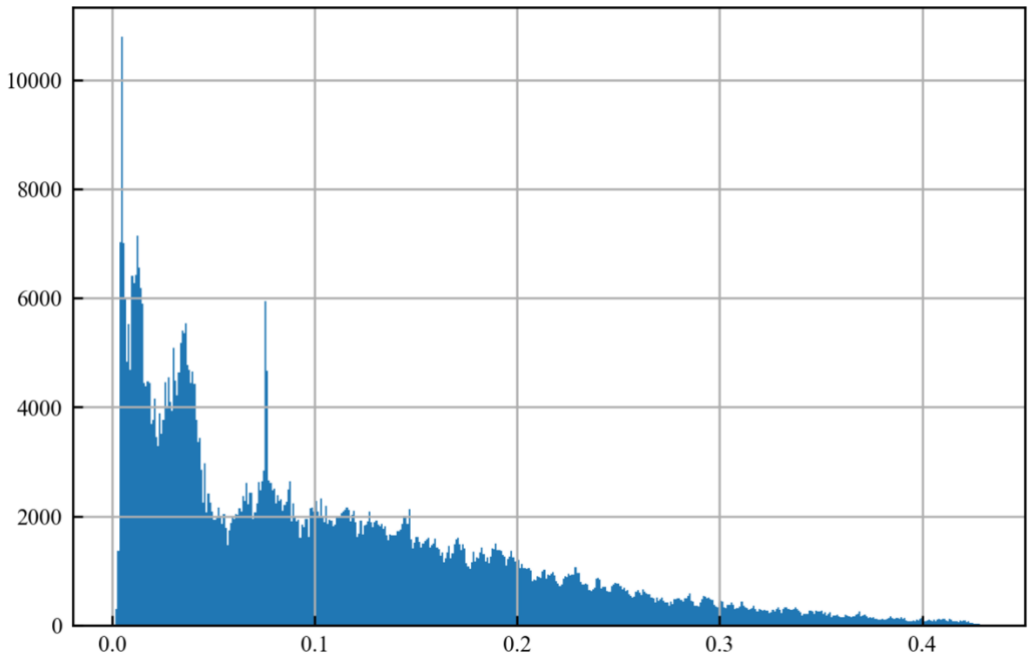

图 5.3 类星体后验概率分布的直方图

我们选取 P_real>0.3 的部分，设为 PredGPQ_30，以尽可能高的阈值来剔除尽可能多的污染源。PredGPQ_30 一共包含 28591 个类星体候选体。画出 PredGPQ_30 在颜色空间的分布，并与 3.1 中的类星体训练样本和 3.3 得到的恒星训练样本比较如下：

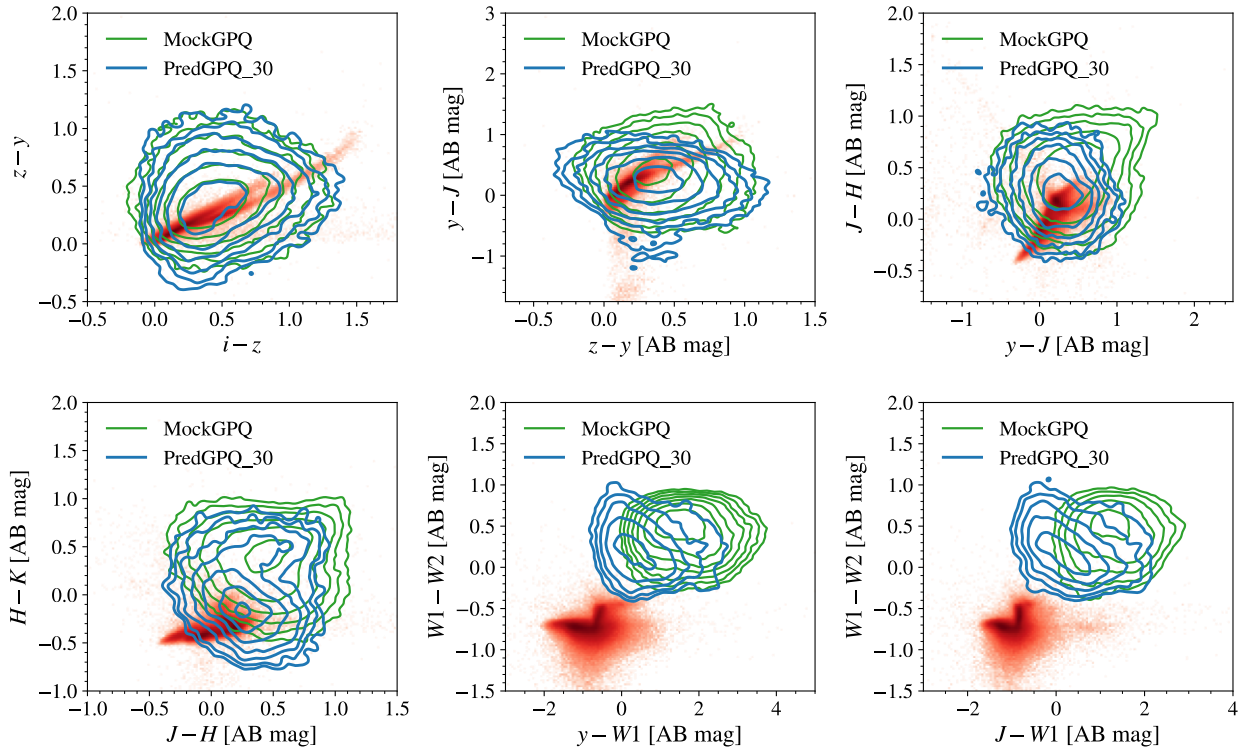

图 5.4 PredGPQ_30 的颜色空间分布

可以看出，相比 PredGPQ，PredGPQ_30 在颜色图上的分布更接近类星体训练样本的分布，红外波段的颜色更远离恒星位点，光学波段颜色在恒星训练样本方向的弥散也减轻了许多。

5.3 利用颜色截断提纯类星体候选体

根据 3.4 中训练样本在颜色空间的分布，我们可通过设置颜色截断直线来去除和恒星样本重叠较大的类星体候选体，进一步提纯 PredGPQ_30。即对 PredGPQ_30 的特征进行以下筛选：i-z<1.5 & y-J>-0.8 & H-K+0.8*(J-H) >0.2 & W1-W2+0.36*(y-W1)>-0.07 & W1-W2+0.33*(J-W1)>0.01。

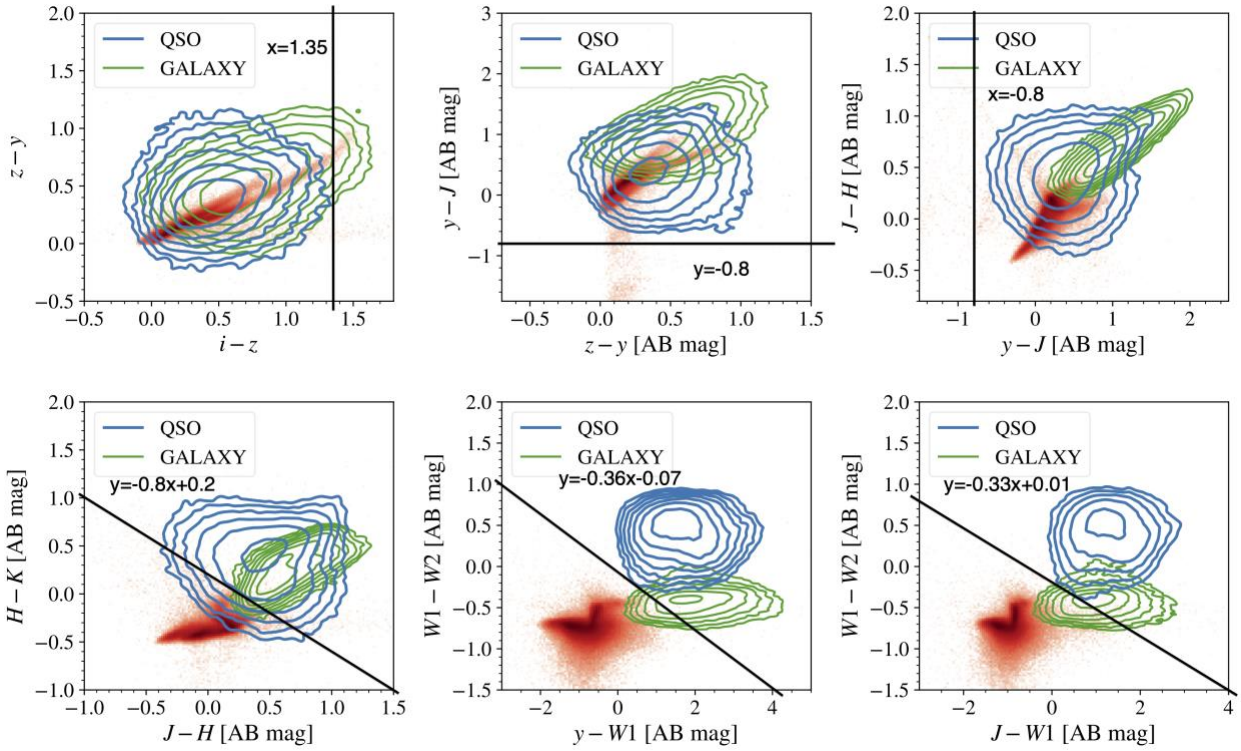

图 5.5 PredGPQ_30 的颜色截断示意图

通过设置以上的颜色截断，我们得到 PredGPQ_30_colorcut，包含 4222 个类星体候选体。画出 PredGPQ_30_colorcut 在颜色空间的分布，并与 3.1 中的类星体训练样本和 3.3 得到的恒星训练样本比较如下：

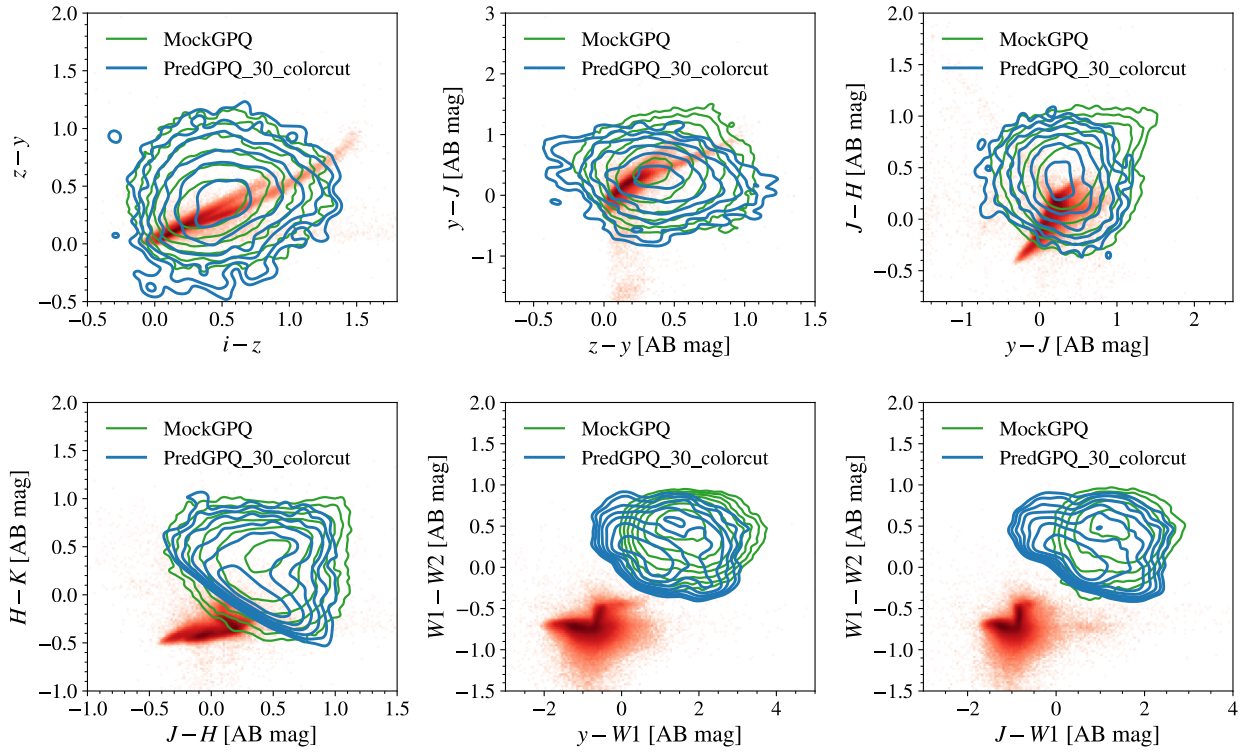

图 5.6 PredGPQ_30_colorcut 的颜色空间分布

可以看出， PredGPQ_30_colorcut 的类星体候选体中恒星污染更少，更加纯净。

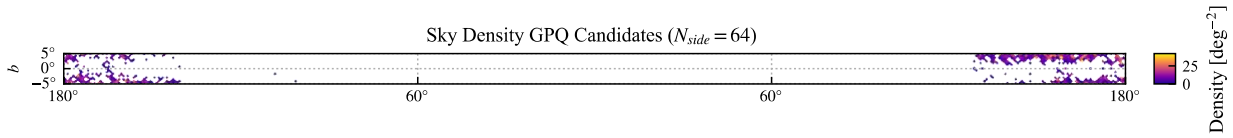

图 5.7 PredGPQ_30_colorcut 的天区密度分布

从上图可以看出，我们最终得到的银纬|b| < 5 deg 的 4222 个类星体候选体主要分布在反银心的区域，纬度越低找到的候选体越少，这也同时反映出银心区域消光高、星场密集的困难所在。

第六章 结论与展望

本文对 Fu et al. (2021)构建的银道面背景类星体选源方法的多个方面进行改进，将其应用到探测难度更大的银纬$|b|<5$ deg 区域，最终在目标天区选出 4222 个类星体候选体。

首先，我们对模拟类星体和星系样本的方法进行了改进，来扩充训练样本的数量。基于多轮模拟的理念，我们进行了两种不同的尝试——无放回随机抽样和有放回随机抽样，并从理论支撑和实际模拟效率两个方面进行比较，最终选择理论性更强、模拟效率更高的有放回随机抽样对高银纬的类星体和星系样本进行十轮模拟，使得高银纬样本的利用率分别提高到 95%和 99%以上。

然后，我们对机器学习应用的分类特征进行修改。由于银纬$|b|<5$ deg 区域星际红化更加严重，我们去掉了消光严重的 PS1 蓝端的 g 和 r 波段，加入了近红外 UKIDSS J,H,K 波段，构建了 i-z, z-y, y-J, J-H, H-K, i-W1, z-W1, y-W1, J-W1, H-W1, K-W1, W1-W2 这 12 个测光颜色特征。

训练出机器学习分类模型后，我们同样通过 Healpixmap 计算出类星体先验概率的分布来校正机器学习预测概率的偏差。但由于原先 Nside=64 的分辨率下天区密度的边界异常值严重，我们将 Nside 调高到 256，并手动矫正异常值，以平滑类星体的先验分布。

在将分类模型应用到目标天区时，我们加严对测试样本质量的筛选，以提高测试样本对目标天区的代表性。基于贝叶斯公式，我们通过先验概率矫正后得到类星体后验概率的分布，并根据后验概率的分布选取大于阈值 0.3 的部分进行颜色截断。通过比较类星体候选体和训练样本在颜色空间的分布，我们设置一系列颜色截断的判据进行筛选，得到更纯净的类星体候选体样本共 4222 个。

上述选出的银纬$|b|<5$ deg 天区类星体候选体可以为 GaiaNIR（Hobbs & Høg, 2018）提供低银纬的天体测量参考架，以提高在银道面中心天区天体测量参数的精度。GaiaNIR 是一项正在计划中的全天巡天任务，旨在改进 Gaia 仅在光学波段探测的局限性，将会提供近红外波段全天的测光数据和天体测量数据，把对天体测量参数的探测拓展到消光更大、红移更高的银道面中心。可见，将银道面背景类星体的选源缩小到银纬$|b|<5$ deg 天区是具有巨大科学价值的工作。

后续的工作将关注于申请望远镜进行类星体候选体的证认，以及对目前的选源方法进行进一步改进。在训练样本的构建方面，我们可以进一步扩充恒星样本的矮星数量，以减小训练样本和实际天区的差异。在训练特征的选择方面，我们可以利用类星体的强 X

射线辐射，加入如 X 射线硬度比、X 射线流量、X 射线-光学流量比、X 射线-红外流量比作为分类特征（参考 Lin et al. 2012），以更好地区分类星体和恒星。在提纯类星体候选样本方面，我们可以加入 Gaia 的自行截断，根据类星体自行为 0 的特征，通过限制自行为 0 的概率密度分布来进行进一步筛选。

参考文献

[1] Antonucci, R. (1993). *Unified models for active galactic nuclei and quasars.* Annual review of astronomy and astrophysics, 31(1), 473-521.

[2] Di Matteo T, Springel V, Hernquist L. *Energy input from quasars regulates the growth and activity of black holes and their host galaxies[J]*. nature, 2005, 433(7026): 604-607.

[3] Kormendy J, Ho L C. *Coevolution (or not) of supermassive black holes and host galaxies[J]*. Annual Review of Astronomy and Astrophysics, 2013, 51: 511-653.

[4] Trump J R, Hall P B, Reichard T A, et al. *A catalog of broad absorption line quasars from the sloan digital sky survey third data release[J]*. The Astrophysical Journal Supplement Series, 2006, 165(1): 1.

[5] Eisenstein D J, Weinberg D H, Agol E, et al. *SDSS-III: Massive spectroscopic surveys of the distant universe, the Milky Way, and extra-solar planetary systems[J]*. The Astronomical Journal, 2011, 142(3): 72.

[6] Blanton M R, Bershady M A, Abolfathi B, et al. *Sloan digital sky survey IV: Mapping the Milky Way, nearby galaxies, and the distant universe[J]*. The Astronomical Journal, 2017, 154(1): 28.

[7] Zhao G B, Wang Y, Saito S, et al. *The clustering of the SDSS-IV extended Baryon Oscillation Spectroscopic Survey DR14 quasar sample: a tomographic measurement of cosmic structure growth and expansion rate based on optimal redshift weights[J]*. Monthly Notices of the Royal Astronomical Society, 2019, 482(3): 3497-3513.

[8] Schmidt M, Green R F. *Quasar evolution derived from the Palomar bright quasar survey and other complete quasar surveys[J]*. Astrophysical Journal, Part 1 (ISSN 0004-637X), vol. 269, June 15, 1983, p. 352-374., 1983, 269: 352-374.

[9] Hewett P C, Foltz C B, Chaffee F H. *The large bright quasar survey. 6: Quasar catalog and survey parameters[J]*. The Astronomical Journal, 1995, 109: 1498-1521.

[10] Croom, S. M., Smith, R. J., Boyle, B. J., Shanks, T., Miller, L., Outram, P. J., & Loaring, N. S. (2004). *The 2dF QSO Redshift Survey–XII. The spectroscopic catalogue and luminosity function*. Monthly Notices of the Royal Astronomical Society, 349(4), 1397- 1418.

[11] Lyke, B. W., Higley, A. N., McLane, J. N., Schurhammer, D. P., Myers, A. D., Ross, A. J., ...

& Weaver, B. A. (2020). *The Sloan Digital Sky Survey Quasar Catalog: Sixteenth Data Release*. The Astrophysical Journal Supplement Series, 250(1), 8.

[12] Grazian A, Cristiani S, D'Odorico V, et al. *The asiago-eso/rass qso survey. i. the catalog and the local qso luminosity function[J]*. The Astronomical Journal, 2000, 119(6): 2540.

[13] Green, R. F., Schmidt, M., & Liebert, J. (1986). *The Palomar-Green catalog of ultraviolet-excess stellar objects*. The Astrophysical Journal Supplement Series, 61, 305-352.

[14] Wu X B, Jia Z. *Quasar candidate selection and photometric redshift estimation based on SDSS and UKIDSS data[J].* Monthly Notices of the Royal Astronomical Society, 2010, 406(3): 1583-1594.

[15] Lacy M, Storrie-Lombardi L J, Sajina A, et al. *Obscured and unobscured active galactic nuclei in the Spitzer Space Telescope First Look Survey[J]*. The Astrophysical Journal Supplement Series, 2004, 154(1): 166.

[16] Mateos S, Alonso-Herrero A, Carrera F J, et al. *Using the Bright Ultrahard XMM–Newton survey to define an IR selection of luminous AGN based on WISE colours[J]*. Monthly Notices of the Royal Astronomical Society, 2012, 426(4): 3271-3281.

[17] Gregg M D, Becker R H, White R L, et al. *The FIRST bright QSO survey[J].* arXiv preprint astro-ph/9604148, 1996.

[18] Becker R H, White R L, Gregg M D, et al. *The FIRST Bright Quasar Survey. III. The South Galactic Cap[J]*. The Astrophysical Journal Supplement Series, 2001, 135(2): 227.

[19] Dobrzycki A, Macri L M, Stanek K Z, et al. *Variability-selected quasars behind the Small Magellanic Cloud[J].* The Astronomical Journal, 2003, 125(3): 1330.

[20] Lawrence A, Warren S J, Almaini O, et al. *The UKIRT infrared deep sky survey (UKIDSS)[J].* Monthly Notices of the Royal Astronomical Society, 2007, 379(4): 1599-1617.

[21] Jin X, Zhang Y, Zhang J, et al. *Efficient selection of quasar candidates based on optical and infrared photometric data using machine learning[J].* Monthly Notices of the Royal Astronomical Society, 2019, 485(4): 4539-4549.

[22] Chambers K C, Magnier E A, Metcalfe N, et al. *The pan-starrs1 surveys[J]*. arXiv preprint arXiv:1612.05560, 2016.

[23] Wright E L, Eisenhardt P R M, Mainzer A K, et al. *The Wide-field Infrared Survey Explorer (WISE): mission description and initial on-orbit performance[J]*. The Astronomical Journal,

2010, 140(6): 1868.

[24] Mainzer A, Bauer J, Grav T, et al. *Preliminary results from NEOWISE: an enhancement to the wide-field infrared survey explorer for solar system science[J].* The Astrophysical Journal, 2011, 731(1): 53.

[25] Bailer-Jones C A L, Fouesneau M, Andrae R. *Quasar and galaxy classification in Gaia Data Release 2[J].* Monthly Notices of the Royal Astronomical Society, 2019, 490(4): 5615-5633.

[26] Gaia C, Brown A G A, Vallenari A, et al. *Gaia Data Release 2 Summary of the contents and survey properties[J]*. Astronomy & Astrophysics, 2018, 616(1).

[27] Ben Bekhti N, Winkel B, Richter P, et al. *An absorption-selected survey of neutral gas in the Milky Way halo. New results based on a large sample of Ca ii, Na i, and H i spectra towards QSOs[J].* Astronomy and Astrophysics, 2012, 542: A110.

[28] Westmeier T. *A new all-sky map of Galactic high-velocity clouds from the 21-cm HI4PI survey[J]*. Monthly Notices of the Royal Astronomical Society, 2018, 474(1): 289-299.

[29] Gaia C, Brown A G A, Vallenari A, et al. *Gaia Data Release 2 Summary of the contents and survey properties[J]*. Astronomy & Astrophysics, 2018, 616(1).

[30] Flesch E W. *The Million Quasars (Milliquas) v7.2 Catalogue, now with VLASS associations. The inclusion of SDSS-DR16Q quasars is detailed[J].* arXiv preprint arXiv:2105.12985, 2021.

[31] Kirkpatrick J D, Henry T J, Irwin M J. *Ultra-cool M dwarfs discovered by QSO surveys. I: the APM objects[J]*. Astronomical Journal v. 113, p. 1421-1428 (1997), 1997, 113: 1421-1428.

[32] Vennes S, Smith R J, Boyle B J, et al. *White dwarfs in the 2dF QSO Redshift Survey–I. Hydrogen-rich (DA) stars[J]*. Monthly Notices of the Royal Astronomical Society, 2002, 335(3): 673-686.

[33] Chiu K, Fan X, Leggett S K, et al. *Seventy-one new l and t dwarfs from the sloan digital sky survey[J].* The Astronomical Journal, 2006, 131(5): 2722.

[34] Im M, Lee I, Cho Y, et al. *Seoul national university bright quasar survey in optical (Snuqso). II. Discovery of 40 bright quasars near the galactic plane[J]*. The Astrophysical Journal, 2007, 664(1): 64.

[35] Skrutskie M F, Cutri R M, Stiening R, et al. *The two micron all sky survey (2MASS)[J]*. The Astronomical Journal, 2006, 131(2): 1163.

[36] Kozłowski S, Kochanek C S. *Discovery of 5000 active galactic nuclei behind the magellanic*

clouds[J]. The Astrophysical Journal, 2009, 701(1): 508.

[37] Stern D, Eisenhardt P, Gorjian V, et al. *Mid-infrared selection of active galaxies[J].* The Astrophysical Journal, 2005, 631(1): 163.

[38] Huo Z Y, Liu X W, Xiang M S, et al. *The LAMOST survey of background quasars in the vicinity of M31 and M33–III. results from the 2013 regular survey[J].* Research in Astronomy and Astrophysics, 2015, 15(8): 1438.

[39] Quinonero-Candela, Joaquin, et al., eds. *Dataset Shift in Machine Learning.* MIT Press, 2022.

[40] Fu, Y., Wu, X. B., Yang, Q., Brown, A. G., Feng, X., Ma, Q., & Li, S. (2021). *Finding Quasars behind the Galactic Plane. I. Candidate Selections with Transfer Learning.* The Astrophysical Journal Supplement Series, 254(1), 6.

[41] Pan S J, Yang Q. *A survey on transfer learning[J].* IEEE Transactions on knowledge and data engineering, 2010, 22(10): 1345-1359.

[42] Wang S, Chen X. *The optical to mid-infrared extinction law based on the APOGEE, Gaia DR2, Pan-STARRS1, SDSS, APASS, 2MASS, and WISE surveys[J].* The Astrophysical Journal, 2019, 877(2): 116.

[43] Marocco F, Eisenhardt P R M, Fowler J W, et al. *The CatWISE2020 Catalog[J].* The Astrophysical Journal Supplement Series, 2021, 253(1): 8.

[44] Vallenari A, Brown A G A, Prusti T. *Gaia Data Release 3. Summary of the content and survey properties[J]*. Astronomy & Astrophysics, 2022.

[45] Creevey O L, Sarro L M, Lobel A, et al. *Gaia Data Release 3: A golden sample of astrophysical parameters[J]*. arXiv preprint arXiv:2206.05870, 2022.

[46] Sarro L M, Berihuete A, Smart R L, et al. *Ultracool dwarfs in Gaia DR3[J]*. arXiv preprint arXiv:2211.03641, 2022.

[47] York D G, Adelman J, Anderson Jr J E, et al. *The sloan digital sky survey: Technical summary[J]*. The Astronomical Journal, 2000, 120(3): 1579.

[48] Wu Q, Shen Y. *A Catalog of Quasar Properties from Sloan Digital Sky Survey Data Release 16[J]*. The Astrophysical Journal Supplement Series, 2022, 263(2): 42.

[49] Abdurro'uf Y T L, Hirashita H, Morishita T, et al. *Dissecting Nearby Galaxies with piXedfit. I. Spatially Resolved Properties of Stars, Dust, and Gas as Revealed by Panchromatic SED Fitting[J]*. The Astrophysical Journal, 2022, 926(1): 81.

[50] Planck Collaboration 2016, Ade P A R, Aghanim N, Arnaud M, et al. *Planck 2015 results-xiii. cosmological parameters[J].* Astronomy & Astrophysics, 2016, 594: A13.

[51] Green G M. dustmaps: *A Python interface for maps of interstellar dust[J].* Journal of Open Source Software, 2018, 3(26): 695.

[52] West A A, Morgan D P, Bochanski J J, et al. *The sloan digital sky survey data release 7 spectroscopic M dwarf catalog. I. Data[J]*. The Astronomical Journal, 2011, 141(3): 97.

[53] Li J, Liu C, Zhang B, et al. *Stellar parameterization of LAMOST M dwarf stars[J].* The Astrophysical Journal Supplement Series, 2021, 253(2): 45.

[54] Wang Y F, Luo A L, Chen W P, et al. *Ultracool dwarfs identified using spectra in LAMOST DR7[J]*. Astronomy & Astrophysics, 2022, 660: A38.

[55] Lodieu N, Contreras M E, Osorio M R Z, et al. *New ultracool subdwarfs identified in large-scale surveys using Virtual Observatory tools-I. UKIDSS LAS DR5 vs. SDSS DR7[J]*. Astronomy & Astrophysics, 2012, 542: A105.

[56] Kirkpatrick J D, Martin E C, Smart R L, et al. *Preliminary trigonometric parallaxes of 184 late-T and Y dwarfs and an analysis of the field substellar mass function into the "planetary" mass regime[J].* The Astrophysical Journal Supplement Series, 2019, 240(2): 19.

[57] Zhang Z H, Galvez-Ortiz M C, Pinfield D J, et al. *Primeval very low-mass stars and brown dwarfs–IV. New L subdwarfs, Gaia astrometry, population properties, and a blue brown dwarf binary[J].* Monthly Notices of the Royal Astronomical Society, 2018, 480(4): 5447-5474.

[58] Chen T, Guestrin C. *Xgboost: A scalable tree boosting system[C]*//Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016: 785-794.

[59] Yi Z, Chen Z, Pan J, et al. *An efficient spectral selection of M giants using XGBoost[J]*. The Astrophysical Journal, 2019, 887(2): 241.

[60] Acero F, Ackermann M, Ajello M, et al. *Fermi large area telescope third source catalog[J]*. The Astrophysical Journal Supplement Series, 2015, 218(2): 23.

[61] Mirabal N, Charles E, Ferrara E C, et al. *3FGL demographics outside the galactic plane using supervised machine learning: Pulsar and dark matter subhalo interpretations[J]*. The Astrophysical Journal, 2016, 825(1): 69.

[62] Akiba T, Sano S, Yanase T, et al. *Optuna: A next-generation hyperparameter optimization*

framework[C]//Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. 2019: 2623-2631.

[63] Chicco D, Jurman G. *Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone[J]*. BMC medical informatics and decision making, 2020, 20(1): 1-16.

[64] Davis J, Goadrich M. *The relationship between Precision-Recall and ROC curves[C]*//Proceedings of the 23rd international conference on Machine learning. 2006: 233-240.

[65] Saito T, Rehmsmeier M. *The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets[J]*. PloS one, 2015, 10(3): e0118432.

[66] Hobbs D, Høg E. *GaiaNIR–A future all-sky astrometry mission[J].* Proceedings of the International Astronomical Union, 2017, 12(S330): 67-70.

Lin, D., Webb, N. A., & Barret, D. (2012). *Classification of X-Ray Sources in the XMM-Newton Serendipitous Source Catalog.* The Astrophysical Journal, 756(1), 27.

致谢

行文至此，感慨万千。无从落笔，总觉词穷不尽意。

感谢北京大学物理学院天文学系的吴学兵教授。吴老师是我的毕业论文指导老师，是我的本科生科研指导老师，也是我大学四年期间的班主任。犹记大二大三的时候，感觉前路一片迷茫，不知所向，更不知自己想要为之奉献的事业究竟是何。是和吴老师的一次次交流，才坚定了我如今努力的方向。大二下学期的时候，为了确定本研的方向，吴老师作为班主任，帮我分析了所里各个组老师的具体课题，最终我选择进入吴老师的组，做机器学习搜寻类星体的相关工作。大四开学之初，因为暑研进展不顺利，对于留学申请多有迷茫，我再次找到吴老师寻求帮助，诉说我对未来就业市场、读博方向、学术热情的苦恼。吴老师点醒了我：不要现在就考虑这么长远的事情，走一步算一步，走一步看一步，不要让对未来的担忧困住了你前进的脚步。吴老师的悉心教导，将指引着我未来所有的上下求索，成为我冲破迷雾寻找自我的勇气。

感谢北京大学科维理天体物理研究所的傅煜铭博士。煜铭学长（曾经）是我们组的大师兄，是组里机器学习类星体搜寻课题的主导者，是我本科生科研和本科生毕业论文这漫漫长路上的灯。学长是我学术路上的引路人，是我科研的启蒙者，可以说是手把手地带领着我从一个彻彻底底的菜鸟、小白走到了现在。记得当年刚进组的时候，最怕的就是回答不上来学长的提问，听不懂学长讲授的知识，搞不懂学长让我做的工作。所以也曾战战兢兢，如履薄冰，组会仿佛成了我最害怕的事情之一，科研仿佛成了一支可能随时走火射中我的枪。转折点在大三上学期的一次小组讨论，学长让我用 Python 写一个做样本循环模拟的封装程序。幸好，大一下学期数据结构与算法给我打下了比较坚实的编程基础，两天内我便成功地写完了这个程序，也得到了学长第一次满意的夸赞。从此以后，我逐渐找回了我在科研中的信心，努力发挥自己的代码强项，尽可能地做好我的本职工作。两年的本研训练中，最令我感动、钦佩的不仅仅是学长仿佛无所不知、无所不能的博学和精深，还有学长恒久不变的耐心、助人为乐的性格、以及对科研和学术的无限热情和忠诚。每次不论遇到什么问题，若再三思考无果，便向学长求助，总能得到直切要害的启发；有了新的科研想法，就跟学长讨论，不论是多么荒谬的想法，学长也会认真思考，给出分析；在别的研究项目中遇到了问题，学长也会尽可能地帮助我解答疑难。可以说，学长是我一辈子学习的楷模，是我今后事业的标杆，也是我一辈子需要心怀感恩的对象。能和煜铭学长合作课题，是我人生中的最佳幸运之一。也希望今后如果还有机会，能够和学长再次合作。

感谢我的暑研导师——澳大利亚国立大学的 Yuan-Sen Ting 教授，以及我的暑研项目合作者——李嘉轩师兄。二位都是在学术领域一等一的大牛，能够与之合作，是我荣幸之至。Prof. Ting 在天文统计学领域的创新性造诣以及对计算机科学领域和天体物理领域前沿交叉的深耕，是我所异常钦佩的地方。我的暑研题目是 Population Inference for Stellar Properties with Neural Density Estimation，其研究的难度、份量和前瞻性之高，极大地拓宽了我的学术视野，同时对我来说也是一次非常严峻的人生考验。由于基础不够强，暑假时间也不够充裕，我的暑研项目并没有完成地很好，但经过这次项目，我的科研能力得到了极高强度的训练。由于疫情，未能到线下进行暑研，不巧 Prof. Ting 也是比较繁忙，无法进行紧密的指导和联系，大部分项目都是在经过我一个人独立而痛苦地沉淀后，一步步移山填海式地完成的。虽然没有做出很好的成果，也面临过无数个凌晨、通宵调试代码的崩溃，但更重要的是，我的科研素养、英语交流能力，以及独立思考能力、自主学习能力、自我内驱力、抗压能力，都得到了极大的提升，这也将使得我受益终生。同时，非常幸运能够在做海外暑研的时候，和北大天文系本科出身的直系学长合作，并得到有关留学申请、学术科研上的指导和建议。嘉轩学长目前是普林高研院 Phd 在读，是天文学生的榜样，从他那里讨教的人生经验，都是我人生中宝贵的财富。也是在嘉轩学长的帮助和指导下，我以愚公移山的信念摸着石头过河，爬过了暑研这段荆棘密布的山路。学长曾经问过我，为什么想要转专业去读统计的 Master。我说，因为我实在不确定对 Astrostat 的热爱究竟是出于对 Statistical and Computer Science 的兴趣，还是只是在舒适圈内的自我麻痹，所以想走出去多看看世界，寻找自己真正的兴趣所在。能够得到学长的理解与鼓励，也是我坚定跨申统计、出国留学的动力之一。

感谢我现生中的朋友们。感谢我的舍友陈湘、文雨涵、石甄，可以容忍我日常生活中的小毛病，陪伴我大学四年的成长。我们共同度过的时光，是我青春画卷中平淡但隽永的一笔。感谢王一川、刘宇飞、范晓燕，有幸初中相识，高中结义，也有幸到了同一座城市读大学。我们构成了彼此的芳华，即使各奔天涯也要成为彼此的港湾。也感谢你，为我编织了那场梦，感谢你做我心灵的治疗师。在那段被泪水打湿的日子里，你是我流淌出的唯一甜蜜的几滴。曾经有一杯粉色的气泡水，但可惜我没有一饮而尽，水里只剩下了糖精。

感谢北京大学和北京大学物理学院，这片孕育我的热土。感谢 2019 年的一二·九演唱会，有幸在众人面前弹响百周年纪念讲堂的施坦威，让星河的隽永也从我的指尖流过。感谢物院羽毛球队的队友们，有幸和你们一起打球、杀球，一起在 2019 年的新生杯和 2023 年的北大杯赛场上共同拼搏，我们的汗水共同交织成团结的力量。感谢物院学生会这个大

家庭，有幸在文艺部工作两年，组织过大大小小四场晚会和多次活动，体验过幕后的场控组也参与过台前的节目组。感谢你们，我的青春如此充实而精彩。

最后，也是最重要的，感谢我的父母。不论是养育之恩，还是对于我出国留学这件事不遗余力的支持，不论是物质上给我坚实的后盾，还是从精神上给我的依靠和鼓舞。很幸运拥有一个并不 push 我的家庭，能够容忍我间歇性的焦躁不安和负面情绪的宣泄，能够以我的开心健康、平安喜乐为最高纲领。在迷茫的时候，在不开心的时候，在无助的时候，在假如是失去了一切的时候，还好父母还在我的身后。虽然饱受过、也正在饱受很多困难的摧残，但这些终将无法打倒我，因为我的父母希望我幸福。这也是我坚定着要好好活下去的，至上的，勇气和信念。

长路漫漫，往事难追。惟有上下求索，反哺恩情。

北京大学学位论文原创性声明和使用授权说明

北京大学学位论文原创性声明和使用授权说明

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。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名：刘行健 日期：2023年 6 月 1 日

学位论文使用授权说明

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

- 按照学校要求提交学位论文的印刷本和电子版本；
- 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
- 学校可以采用影印、缩印、数字化或其它复制手段保存论文；
- 因某种特殊原因需要延迟发布学位论文电子版，授权学校□一年/□两年/□三年以后，在校园网上全文发布。

（保密论文在解密后遵守此规定）

论文作者签名：刘行健 导师签名：

日期：2023 年 6 月 1 日