中国科学院大学

University of Chinese Academy of Sciences

博士学位论文

基于大型巡天数据的类星体选源及测光红移技术研究

作者姓名：李长华

指导教师：崔辰州 研究员 中国科学院国家天文台

张彦霞 研究员 中国科学院国家天文台

学位类别：理学博士

学科专业：天文技术与方法

培养单位：中国科学院国家天文台

2022 年 12 月

Research on Quasar Candidate Selection and Photometric Redshift Estimation Based on Large-scale Survey

A dissertation submitted to

University of Chinese Academy of Sciences

in partial fulfillment of the requirement

for the degree of

Doctor of Philosophy

in Astronomical Technology

By

Changhua Li

Supervisor: Prof. Chenzhou Cui, Prof. Yanxia Zhang

National Astronomical Observatories, Chinese Academy of Sciences

December, 2022

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：　　导师签名：

日　　期：　　日　　期：

摘 要

类星体是宇宙中一种明亮而遥远的天体，被称为“宇宙探针”，对宇宙学研究具有重要意义。获取大量的类星体样本，尤其是高红移类星体样本有助于对这些研究工作更好地开展及深入。但是，由于类星体观测上具有类似恒星的特征，这无疑加大了发现与识别类星体的难度。为了提高光谱望远镜发现与认证类星体的效率，基于现有的已知样本，应用机器学习方法可以从大型测光巡天望远镜获取的大规模测光数据中搜寻更多高置信度的类星体候选源。

天体的物理性质研究首先要知道天体的距离，精确的距离测量得益于光谱的获得。考虑到光谱观测费时费力且代价高，获得所有天体的光谱显然是不可能的，尤其是暗弱天体。因此，估测天体的测光红移就显得尤为重要。以 SDSS、LAMOST 为代表的大型光谱巡天项目为我们提供了大量的已知样本，利用机器学习和模板匹配方法，设计不同方案开展测光红移的测量工作，对大型巡天的类星体和星系给出红移估测是可行的。

随着大规模巡天望远镜的投入使用，天文学进入到大数据时代。例如，北京-亚历桑那巡天项目（BASS），于 2019 年发布了第三版数据，包括了近 2 亿天体的测光信息。同时，BASS 巡天又是暗能量巡天（DESI）图像巡天的重要组成部分。DESI 于 2019 年发布了第九版图像巡天数据，包括了近 20 亿的天体信息。这些大规模的巡天科学数据为新的科学发现（如高红移类星体、高红移星系）提供了更多的机遇与可能，也带来了前所未有的挑战。这种挑战不仅体现在数据的传输、存储、计算等基础设施，更体现在数据挖掘算法及适应大数据科学的新型科研模式上。在基于大型巡天数据的类星体选源及测光红移研究中，在线科研平台和机器学习应用成为一种必然的选择与发展趋势。

本论文的主要研究工作包括四部分：第一，应用机器学习方法设计了二元分类器和多元分类器。利用这些分类器从 BASS DR3 测光星表中寻找类星体侯选体，为大型光谱巡天提供类星体侯选源。第二，类星体测光红移的测量。比较了不同机器学习方法的性能，基于性能最优的 CatBoost 方法设计了一步模型和两步模型，提高了红移测量精度，完成了 BASS DR3 类星体候选体的红移估测，为挑选高红移类星体样本提供了参考。第三，星系测光红移的测量。比较和应用了机器学习和模板匹配方法在测光红移上的性能和优缺点。指出低红移部分机器学习方法更准确，高红移部分模板匹配结果更可靠，并对 DESI 图像巡天数据 DR9 的星系进行了红移估测，有助于高红移星系的挑选和研究。第四，面向类星体选源与测光红移估计的在线科研平台的探讨与开发。结合在线科研平台，设计了数据交叉和模型应用的并行化处理算法，有助于提升天文学家的大数据处理能力和科研效率。

基于 BASS、SDSS、LAMOST 和 ALLWISE 巡天数据生成的天体分类训练样本，利用 XGBoost 算法在多种特征组合及不同模型参数下进行训练，构建了二元

及多元分类器。多元分类器可以直接将星系、恒星、类星体进行一次区分。在只有光学特征的情况下，最优的准确度为 94.47%；当考虑红外特征情况下，最优的准确度达到了 98.43%。二元分类器先进行点源与展源的区分，再将点源分为恒星与类星体。在只有光学特征的情况下，点展源分类的最优准确度为 97.28%, 恒星与类星体分类的最优准确度为 93.22%；考虑红外特征情况下，点展源分类的最优准确度为 98.67%，恒星与类星体分类的最优准确度达到了 99.15%。然后用构建的最优模型对 BASS DR3 中观测源进行了分类，两种分类情况下的置信度都超过 95% 的类星体侯选体有 798, 928 个。这些类星体候选体可以作为 LAMOST、DESI 或其他巡天计划的输入星表，用于后续跟踪观测。这些分类结果对于 BASS 源的进一步分析和研究具有重要的意义和参考价值。

基于同样的已知类星体训练样本，我们进行了类星体测光红移技术的研究，设计了一步模型和二步模型两种类星体红移的估计方案。在每一种方案中，采用 XGBoost、CatBoost、随机森林三种算法分别进行训练。通过对比，CatBoost 算法的性能优于 XGBoost 和随机森林。因此，CatBoost 作为核心算法，在考虑红外特征的情况下，一步模型的最优性能为 $MSE = 0.1059$，$\sigma_{NMAD} = 0.0872$；两步模型的最优性能为 $MSE = 0.0970$, $\sigma_{NMAD} = 0.0854$。通过对比，两步模型要优于一步模型，尤其在对高红移类星体的预测上。基于两种模型，我们分别对 BASS DR3 的类星体侯选体进行了红移预测，其中红移大于 3.5 的侯选体有 3,938 个，红移大于 4.5 的类星体有 121 个。这些预测结果对类星体的统计研究和高红移类星体证认具有重要的研究价值。

基于 DESI、SDSS、LAMOST 的巡天数据，我们采用 EAZY 和 CatBoost 两种方法进行了红移预测研究。利用光学和红外波段信息，CatBoost 在训练集上获得的最佳性能为 MSE=0.0032，σ_{NMAD}=0.0156，O=0.88%。在具有足够多的已知样本和已知红移范围内，CatBoost 优于 EAZY 方法，但 EAZY 方法更适合于暗源，有助于发现高红移星系。我们利用 CatBoost 和 EAZY 完成了所有 DESI DR9 星系的红移估计，这些数据有助于星系距离的测定及其性质的进一步研究。

在利用机器学习的类星体选源或红移估计的研究中，除算法选择与性能优化外，数据准备与模型应用也是其中的两个重要环节。大规模巡天数据为这两个任务的开展带来了巨大的挑战。在本研究中，我们结合国家天文科学数据中心的计算资源，开展了在线平台与并行化计算技术的研究。通过设计集群环境下的数据交叉、模型应用的并行计算方案，提升了数据处理的效率，为后续进一步建立在线科研服务平台打下了基础。天文学家借助机器学习、人工智能、高性能计算、在线科研平台等众多工具或技术将会做出更多、更好的科学成果。

关键词：天文数据库，类星体，测光红移，机器学习，分类与回归

Abstract

Quasars are the brightest and most distant objects discovered in the Universe. They are essential probes of distant and early Universe and of great significance to cosmological research. Obtaining a large number of quasar samples, especially high redshift quasar samples, will help to better carry out and deepen these researches. However, due to the characteristics of quasars similar to stars, it is undoubtedly more difficult to find and identify quasars. In order to improve the efficiency of spectral telescopes in discovering and identifying quasars, based on the existing known samples, machine learning methods can be used to search for more quasar candidates with high confidence from the large-scale photometric data obtained by the large photometric survey telescopes.

The study of physical properties of celestial bodies should first know their distance. Accurate distance measurement benefits from the acquisition of spectra. It is hard, time-consuming and high-cost work to obtain spectroscopic observation for a large volume of sources. Moreover it is obviously impossible to obtain the spectra of all celestial bodies, especially for faint sources. In this way, it is particularly important to obtain the photometric redshifts of celestial bodies. The large-scale spectroscopic sky survey projects (e.g. SDSS, LAMOST) have provided us with enough known samples. By means of machine learning and template matching methods, different schemes have been designed to carry out the measurement of photometric redshifts. It is feasible to estimate the redshifts of quasars and galaxies in large-scale sky surveys.

With the application of large-scale sky survey telescopes, astronomy has entered the era of big data. For example, the Beijing-Arizona Sky Survey (BASS) Data Release 3 (DR3) catalogue was released in 2019, which contains the data from all BASS and the Mosaic z-band Legacy Survey (MzLS) observations during 2015 January and 2019 March, including about 200 million sources. Meanwhile, BASS sky survey is an important part of the Dark Energy Spectroscopic Instrument (DESI) image survey. DESI released the ninth edition of data in 2019, including nearly 2 billion celestial body information. Large scale scientific data provide more opportunities and possibilities for further astronomical research and new scientific discoveries (e.g. high redshift quasars, high redshift galaxies), as well as unprecedented challenges. This challenge is not only reflected in data storage, computing and other infrastructures, but also in data mining algorithms and new scientific research paradigm that adapt to big data science. Online research and machine learning applications have become an inevitable choice and development trend.

The main research work of this paper consists of four parts: first, the application of machine learning methods to design a binary classifier and a multiclass classifier to

find quasar candidates from the BASS DR3 photometric catalog. The quasar candidates for large spectral surveys are provided. Second, the photometric redshift measurement of quasars. The performance of different machine learning methods is compared, and a one-step model and a two-step model are designed based on the CatBoost method with the best performance. The redshift measurement accuracy is improved, and then the redshift estimation of BASS DR3 quasar candidates is completed, which is helpful for the selection of high redshift quasar cnadidates. Third, the photometric redshift measurements of galaxies. The performance as well as advantages and disadvantages of machine learning and template matching methods on redshift estimation is compared and applied. It is found that the machine learning method is more accurate for low redshift part, while the template matching is more reliable for high redshift part. The redshift estimation of DESI DR9 galaxies is given, which contributes to the selection and study of high redshift galaxies. Fourth, the exploration and development of an online scientific research platform for quasar candidate selection and photometric redshift estimation. Based on the online scientific research platform, parallelized data processing algorithms are designed, especially for multiwavelength data integration and machine learning application, which is helpful to improve big data processing ability and scientific research efficiency of astronomers.

Based on BASS, SDSS, LAMOST and ALLWISE databases, we first analyze the data characteristics of the training samples, and then trained the XGBoost algorithm with a variety of feature combinations and different model parameters to build binary and multiclass classifiers. The multiclass classifier may directly distinguish galaxies, stars and quasars at one time. In the case of only optical features for the multiclass classifier, the optimal accuracy is 94.47%; when adding infrared features, the optimal accuracy reaches 98.43%. The binary classifier firstly distinguishes point sources from extended sources, and then divides point sources into stars and quasars. Only with optical features for the binary classifier, the optimal classification accuracy of point and extended sources amounts to 97.28%, and the optimal classification accuracy of stars and quasars is 93.22%; considering infrared features, the optimal classification accuracy of point and extended sources arrives at 98.67%, and the optimal classification accuracy of stars and quasars is 99.15%. In brief, the accuracy of these classifiers with the best input patterns is larger than 90.0%. Finally, all selected sources in the BASS DR3 catalogue are classified by these classifiers. The classification label and probabilities for individual sources are assigned by different classifiers. When the predicted results by binary classifier are the same as multiclass classifier with optical and infrared information, the number of quasar candidates reaches 798,928 ($P_{QSO} > 0.95$). Those candidates may be taken as input catalogue of LAMOST, DESI, or other projects for follow-up observation. The classified result will be of great help and reference for future research of the BASS DR3 sources.

Then, we study the photometric redshift measurement technologies of quasars on the same training samples of quasars. XGBoost, CatBoost and random forest are used to build regression models, and the process of searching for optimal features and optimal hyper parameters by grid search is optimized. We design two schemes: one scheme (namely one-step model) is to predict photometric redshifts directly based on the optimal models created by those three algorithms; the other scheme (namely two-step model) is to first classify the data into low- and high-redshift data sets, and then predict photometric redshifts of these two data sets separately. Among the experiments, the performance of CatBoost is better than XGBoost and random forest. Therefore considering optimal and infrared features, CatBoost, as the core algorithm, has the optimal performance of $MSE = 0.1059$, $\sigma_{NMAD} = 0.0872$ for the one-step model; the optimal performance of the two-step model is $MSE = 0.0970$, $\sigma_{NMAD} = 0.0854$. By comparison, the two-step model is better than the one-step model, especially in the prediction of high redshift quasars. Based on the two kinds of models, we predict the photometric redshifts of the quasar candidates of BASS DR3, including 3,938 candidates with redshift $\geq$ 3.5 and 121 quasars with redshift $\geq$ 4.5. These prediction results are of great significance to the statistical study of quasars and the identification of high redshift quasars.

Based on DESI, SDSS, LAMOST and ALLWISE databases, we apply EAZY and CatBoost methods to predict the photometric redshifts of galaxies in the DESI DR9 catalogue. Using optical and infrared information, CatBoost achieves the best performance on the training set with MSE=0.0032, σ_{NMAD}=0.0156, O=0.88%. CatBoost is superior to EAZY in the range of known redshift and with enough known samples, but EAZY method is more suitable for faint sources, which is helpful to discover high redshift galaxies. By means of CatBoost and EAZY, we have completed the photometric redshift estimation of all DESI DR9 galaxies. These data contribute to the distance measurement of galaxies and further study of their properties.

In addition, in order to deal with the challenges brought by the migration and calculation of large-scale sky survey data, based on the computing resources of the National Astronomical Data Center, this thesis carries out the research on the online scientific platform, and designs parallel schemes for different stages of data processing, especially the cross fusion scheme for large-scale catalog data, which improves the efficiency of data processing and lays a foundation for further establishing large-scale data cross fusion online service platform. Astronomers will make more and better scientific results with the help of machine learning, artificial intelligence, high-performance computing, online scientific platform and many other tools or technologies.

Key Words: Astronomical databases, Quasars, Photometric redshifts, Machine learning, Classification and regression

目 录

图目录

表目录

第 1 章　引言

1.1　类星体

1.1.1　类星体的发现

类星体是宇宙中一种明亮而遥远的天体，与脉冲星、星际物质与宇宙背景辐射，并称为 20 世纪 60 年代天文学的“四大发现”。类星体是一种特殊的活动星系核，按照活动星系核统一模型，其中心是一个超大质量的黑洞，周围被吸积盘环绕，当黑洞从吸积盘中吸取物质时，能量以电磁波的形式辐射出去，覆盖了从射电到 γ 射线的全电磁波段。如图1-1所示 [1]，类星体核区内部有剧烈的物理过程而发出强烈的辐射，由于距离遥远使得人们只能看到其格外明亮的核区部分。从可见光图像上，如图1-2左图所示，酷似点源，这跟恒星非常相似，故单从可见光图像区分二者比较困难。图 1-2 右图是 3C 273 类星体的光谱。通常类星体具有强发射线、高光度、大红移等典型特征。类星体比正常星系亮百倍，甚至千倍，它的能量从何而来？如果能够解答这个问题，有可能揭示更多宇宙的奥秘。

图 1-1 类星体的艺术照片。

Figure 1-1 Artist’s impression of a quasar.

1963 年，著名的英国《自然》（Nature）杂志刊出了美籍荷兰天文学家施密特（M.Schmidt）的论文《3C 273: A STAR-LIKE OBJECT WITH LARGE RED-SHIFT》(Schmidt, 1963)。3C 273 是第一个光谱证认的类星体，标志着类星体发现的开端。随着巡天时代的到来，类星体发现的数量急剧上升。例如，帕洛马-格林亮类星

[1] https://cdn.spacetelescope.org/archives/images/screen/opo2010a.jpg

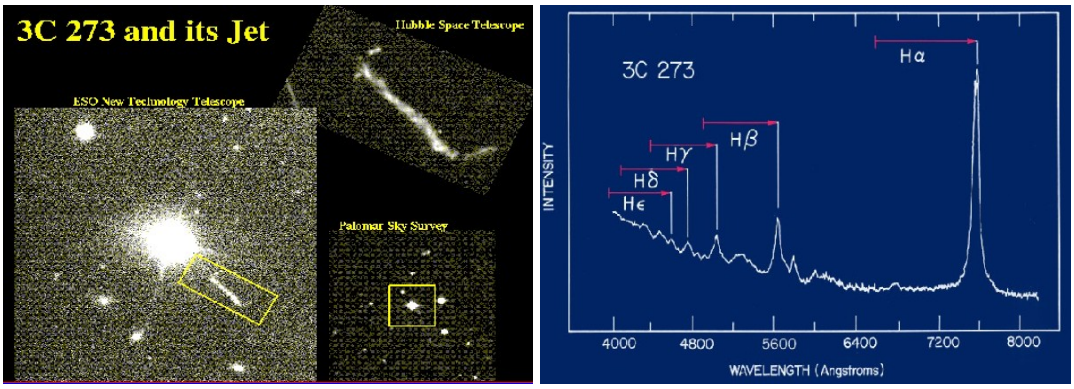


图 1-2 3C 273 的图像及光谱。

Figure 1-2 The image and spectrum of 3C 273.

体巡天（The Palomar-Green Bright Quasar Survey，BQS）(Schmidt et al., 1983) 发现了 114 颗类星体；大型亮类星体巡天（The Large Bright Quasar Survey，LBQS）(Foltz et al., 1987) 证认了 1,055 颗类星体；两度视场类星体红移巡天（The 2dF Quasar Redshift Survey，2QZ）(Boyle et al., 2000) 找到了 23,338 颗类星体；斯隆数字巡天（The Sloan Digital Sky Survey，SDSS）DR16 证认了 750,414 颗类星体；大天区面积多目标光纤光谱天文望远镜（The Large Sky Area Multi-Object Fiber Spectroscopic Telescope，LAMOST，也称郭守敬望远镜）DR8 观测到了 72,061 颗类星体。已有和后续新的巡天项目的发展，为类星体的发现提供了新的契机。

类星体在天文学研究中具有重要的地位和研究价值。通过研究类星体可以研究吸积盘、黑洞的形成和演化、黑洞自旋、双黑洞、黑洞与寄主星系的关系等 (Kormendy et al., 2013)；类星体的吸收线可以作为研究不同红移处的星际及星系际介质的探针；大样本的类星体可以研究宇宙大尺度结构、重子声学震荡；高红移类星体有助于研究宇宙早期的形成和演化 (Blanton et al., 2017)、宇宙早期再电离 (Jiang et al., 2022)、星系的形成与演化等。正因为类星体对宇宙学研究的重要意义，它被天文学家称为“宇宙的探针”。

1.1.2 类星体的预选源方法

由于类星体研究的重要意义，从第一个类星体被发现以来，天文学家一直致力于发现更多的类星体。虽然类星体的最终确认需要拍摄相应的光谱信息，但是进行光谱拍摄时，我们需要提前确认可能发现类星体的坐标位置，以期最大概率的拍到类星体的光谱。这种为光谱望远镜提供类星体侯选体的过程称为类星体预选源。预选源方法主要包括紫外超、多色测光、光变、多波段交叉证认、机器学习等方法。

(1) 光学波段

依据类星体的强发射线及极宽的连续谱方面的特征，在光学波段，类星体与恒星、星系具有不同的颜色。因此，许多类星体巡天喜欢采用颜色标准来挑选类星体侯选源，而且颜色预选源相对来说比较容易操作。设计的巡天望远镜只要拍

下同一天区至少两个波段的图像，就可以通过颜色进行辩认。但是，并不是所有颜色都具有明显的辨别能力。根据类星体的幂律连续谱特征，表明类星体具有明显的紫外超。因此，测量 U、B 波段的星等就可以根据紫外超进行预选源。最早的帕洛马-格林亮类星体巡天就是采用这种方法选取类星体候选体的，但成功率只有 7%。1980 年开展的大型亮类星体巡天也采用了这种方法。但是不同天区，判据也有差别。为了提高颜色选择类星体被证认的概率，减少恒星污染，人们开始用多色方法来进行选源。两度视场类星体红移巡天应用 U、B_J、R 三个波段的星等，在（U-B_J）–（B_J - R）二维空间上选择类星体侯选源。基于 SDSS 巡天的 $ugriz$ 五个光学波段，多色方法也成为 SDSS 巡天选源的主要方法之一，多色方法的证认概率约为 10%。由于测光星等受到星际物质及灰尘吸收的干扰，使得类星体的观测特征受到影响，从而增加了选源的难度。

类星体在光学波段的特征还有比如光变特征，类星体的光变时标为年，但变幅较小；此外，类星体是遥远的天体，其自行为 0，这些特征都可以作为类星体选择的标准 (Hawkins, 1983; Meusinger et al., 2002)。

(2) 射电波段

虽然最早发现的类星体都是射电类星体，但实际上射电类类星体约占类星体总数的 10% 左右。由于射电巡天不受红移影响，通过射电巡天有可能发现高红移类星体。FBQS（FIRST Bright Quasar Survey)(Gregg et al., 1996; White et al., 2000; Becker et al., 2001) 证认类星体的效率较高，超过了 60%，发现了近千颗类星体。

(3) 红外波段

根据类星体具有较强的红外辐射的特点，我们可以开展红外波段与光学波段相结合的方法来搜寻类星体。但是受大气与环境影响，红外波段一般只能利用卫星进行观测。典型代表有 IRAS 巡天 (Soifer et al., 1986)、2MASS 巡天 (Kleinmann, 1992) 及 WISE 巡天 (Wright et al., 2010)。正常星系的红外流量在 60-100μ 波段，但类星体的的红外流量通常处在 12-25μ 波段。因此，在一些光学巡天中加入红外波段及颜色信息，可以提高类星体的证认效率，同时也有助于发现高红移类星体。

(4) γ 射线波段

γ 射线辐射是类星体区别于恒星的一个显著特征。类星体有 γ 射线辐射，而普通恒星则不存在 γ 射线辐射。但是目前在寻找 γ 射线源的技术上还不成熟，探测器的分辨率不高。康普敦 γ 射线天文台的 EGRET 望远镜（the Energetic Gamma-Ray Experiment Telescope），覆盖全天，可以探测到 30MeV-3GeV 范围内的辐射能量，发现了不少类星体 (Thompson et al., 1995)。

(5) X 射线波段

X 射线也是类星体的主要观测特征，并且流量与可见光波段相当，而只有少量的恒星与星系具有 X 射线辐射。因此，X 射线巡天也是一个有效搜寻类星体的方法。EMSS（the Einstein Medium Sensitivity Survey）巡天仅覆盖了 780 平方度的天区，大约发现了 1,400 个源，而其中 30% 是类星体与活动星系核，这表明

了 X 射线巡天对于发现类星体的高效性。

(6) 多波段巡天与机器学习

利用不同波段的信息进行综合考虑，可以有效提高类星体的认证效率。何香涛等 (He et al., 2001) 结合 X 射线波段与射电波段的信息，在选取的 30 个候选体中，3 个是已知的类星体，7 个是已在北京天文台望远镜证认的活动星系核，其它 20 个进行光谱观测，证认了 12 个活动星系核，获得了较高的成功率。SDSS 巡天包括了 *ugriz* 五个波段的信息，许多天文学家结合 SDSS 的测光及 2MASS、WISE，UKIDSS 等红外巡天相结合进行类星体候选体的选择 (Wu et al., 2010, 2012, 2013)，以及利用分类方法来选择类星体候选体 (Gao et al., 2008; Schindler et al., 2017; Clarke et al., 2020)，为 SDSS 和 LAMOST 光谱巡天提供候选源，大幅增加了类星体发现的概率，截止最新发布的数据 SDSS DR16Q，共包括近 75 万类星体，成为目前最大的类星体样本库，也有使用聚类的学习方法从多波段数据中选择候选体 (Zhang et al., 2004)。因此，随着人们积累的多波段数据规模的不断增大及发现的类星体数量的增加，机器学习方法在类星体选源工作中也越来越受到天文学家的重视，成为大数据时代类星体选源的主流技术。

历经几十年的类星体巡天，已知类星体样本初具规模，成绩显著。然而，根据预测，宇宙中 *r* 波段星等小于 23 的类星体数量约为每平方度 304 个 (Palanque-Delabrouille et al., 2016)，比 23 等更暗的类星体将有更多，因此，已发现的类星体只是宇宙中实际类星体数量的很小一部分，而且分布天区很不均衡。图1-3展示了 SDSS DR16Q 类星体的空间分布；图1-4展示了 SDSS DR6Q 类星体的红移分布；图1-5展示了 SDSS DR6Q 类星体的 *r* 星等分布。由图1-3可以发现，已发现类星体的空间分布很不均匀，大部分分布在高银纬，而且受限于 SDSS 的可观测天区限制，仍然有很多天区目前还未观测；由图1-4可知，已发现的类星体样本大部分集中在低红移区域，红移大于 3 的高红移样本很少；由图1-5显示，类星体的 *r* 星等分布在 15 等到 23 等之间，峰值大约在 21 等。通常红移越高，类星体越暗，暗弱的类星体对于研究银河系之外的黑洞与寄主星系关系、黑洞与核球共动演化、宇宙早期演化等具有更重要的意义。由这三个图可见，暗的、低银纬的和高红移的类星体还有待我们去挖掘发现。低银纬的类星体的发现受限于银河系消光、恒星污染等因素；而暗弱的、高红移类星体的搜寻有赖于望远镜观测技术的提高、选源方法的优化等。考虑到望远镜获取光谱的昂贵性，基于现有的观测手段和观测数据（如 PanSTARRS、GAIA、WISE 等），如何自动地、高效地、准确地选取类星体候选体仍是目前天文学家关注的课题，利用当前流行的机器学习和深度学习等方法解决此课题势在必行。由于天文数据的复杂性、海量性、多波段性、异构性、多源性、时域性，为充分发挥大型设备的科学潜力，对即将运行和在建的项目（如 LSST、CSST 等）的提前预研究也迫在眉睫。

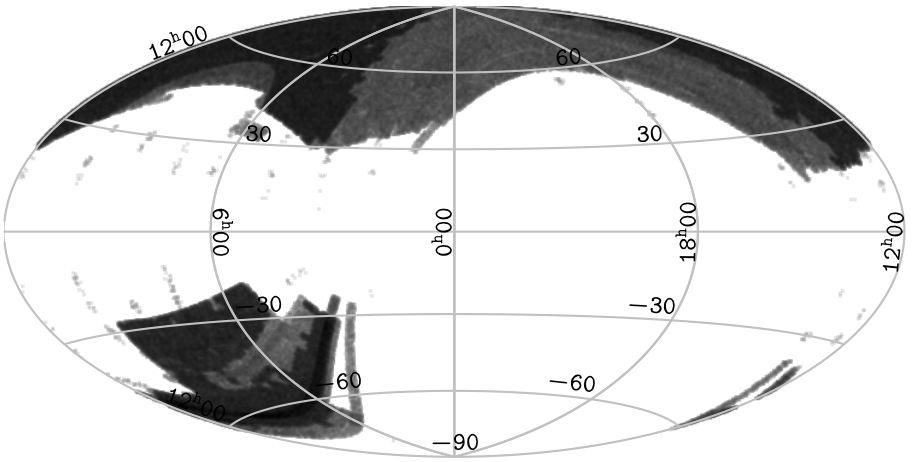


图 1-3 SDSS DR16Q 类星体的银道坐标空间密度分布。

Figure 1-3 The spatial density of SDSS DR16Q Quasars in Galactic Coordinates.

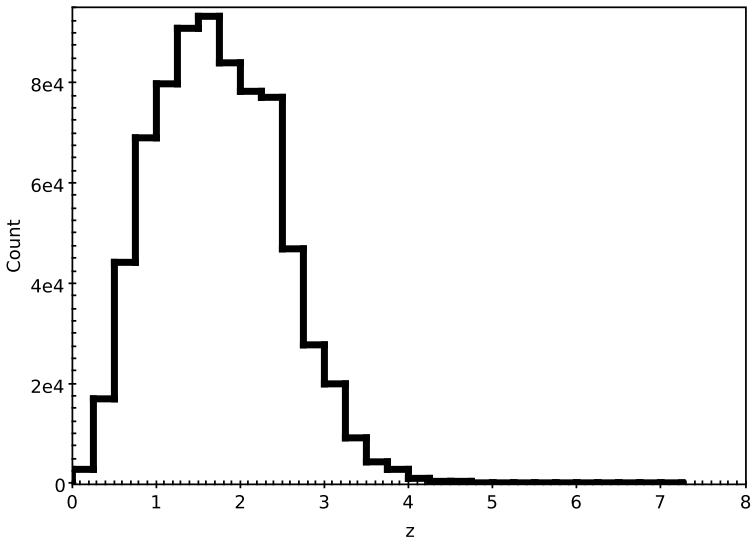


图 1-4 SDSS DR16Q 类星体红移区间分布。

Figure 1-4 The redshift distribution of SDSS DR16Q quasars.

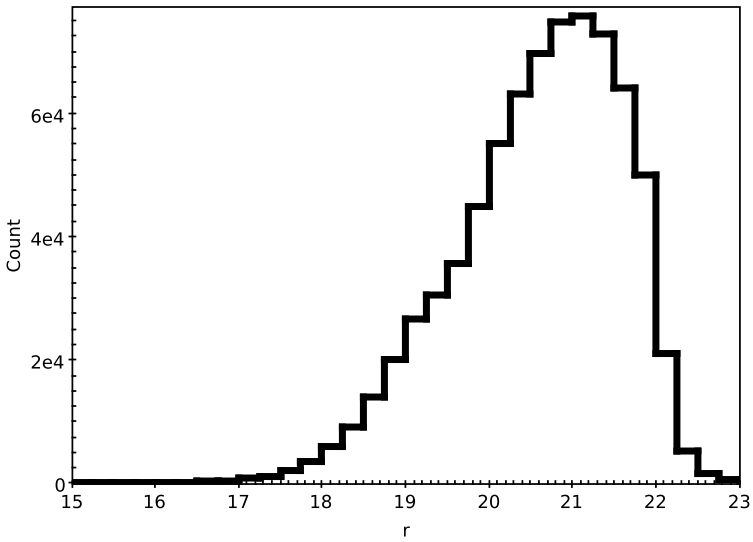


图 1-5 SDSS DR16Q 类星体 *r* 波段星等分布。

Figure 1-5 The *r* magnitude distribution of SDSS DR16Q quasars.

1.2 测光红移

1.2.1 测光红移的概念

当光源远离观测者时，观测者接收到的光波频率比其固有频率低，即向红端偏移，这种现象称为“红移”；而当光源接近观测者时，接收到的频率会比实际频率高，相当于向蓝端偏移，称为“蓝移”。由于宇宙的膨胀，天体相对于观测者始终在以一定的速度退行，因此，天体具有红移特征。每一种谱线的固有波长可以通过静止光源在实验室中进行测量，因此，根据观测天体的同一条谱线的波长，就可以计算出该天体的红移。红移 z 的定义如下：

$$z = \frac{\lambda - \lambda_0}{\lambda_0} \tag{1-1}$$

式中，λ_0 是光源在实验室的某条谱线的静止波长，λ 是观测天体的同一条谱线的波长。1929 年，美国天文学家哈勃经过多年的观测确认，遥远的星系均远离我们地球所在的银河系而去，同时，它们的红移随距离增大而成正比地增加，这一规律称为哈勃定律。哈勃定律的伟大意义，不仅在于它证实了宇宙的膨胀，而且还提供了一种估计宇宙年龄及计算天体距离的方法。根据哈勃定律，对于小红移的天体，距离与红移的关系：

$$d = \frac{cz}{H_0} \tag{1-2}$$

其中 c 是光速，H_0 是哈勃常数 $(75km \cdot s^{-1} \cdot Mpc^{-1})$, z 是红移。而对于大红移的天体，距离与红移的关系为：

$$d_L = \frac{c}{H_0 q_0^2}\left\{q_0 z + (q_0 - 1)[(1 + 2q_0 z)^{\frac{1}{2}} - 1]\right\} \tag{1-3}$$

其中 q_0 是减速因子，d_L 是光度距离。因此，通过红移，我们就可以测量遥远天体的距离。天体距离的测定，对于研究天体的空间位置、形成与演化、求得天体

的光度函数等均具有重要的意义。根据定义，红移的计算需要首先确定天体的光谱谱线的静止波长和位移波长。然而，相比测光观测而言，光谱观测是一种“昂贵”的观测手段，费时费力。到目前为止，人类总共获取的光谱数量还在千万量级，而宇宙中星系的数量处在千亿量级。此外，对于大量暗弱的天体，现有的光谱设备无法获取到有效的光谱信息，因此，天文学家提出基于测光数据来估计天体的红移，这种通过使用中波段、宽波段的测光数据或者图像得到的红移，称为测光红移。测光红移技术的应用有助于快速获取大规模天体的红移数据，为后续的天文学研究提供基础。

1.2.2　测光红移技术

1957 年，Baum 提出利用测光数据来测红移的方法 (Baum, 1957)，并于 1962 年发展了一种估计测光红移的算法 (Baum, 1962)。算法使用光电光度计和 9 个滤光片，覆盖从 3730Å 到 9875Å 的波长范围。首先，他获得了 6 个比较亮的位于室女星系团（Virgo）的椭圆星系的光谱能量分布；随后又得到了 C1095+2044 星系团（又称为 Abell 0801）中的三个椭圆星系的光谱能量分布；最后，他将平均的室女星系团的光谱能量分布与平均的 C10925+2044 星系团的光谱能量分布画在同一张图上进行比较，算出两个能量分布之间的位移，从而可以获得 C1095+2044 星系团的红移 $z = 0.19$。这个结果与光谱红移 $z = 0.192$ 十分接近，由此表明，Baum 的算法是有效的。但其极大依赖于 4000Å 处的光谱截断特征来预测红移，所以只适合用于椭圆星系。1982 年，Puschell 等 (Puschell et al., 1982) 提出利用宽带测光数据预测暗射电星系红移，正式提出测光红移的概念。在 1986 年，Loh 和 Spillar(Loh et al., 1986) 第一次在文章题目中使用了“测光红移”一词，此后，测光红移技术研究得到广泛关注与快速发展。

到了 21 世纪初，模板匹配方法受到天文学家的欢迎 (Wu et al., 2004)，开发了许多模板匹配的软件工具，包括 LePHARE(Arnouts et al., 1999)、BPZ(Benitez, 2000)、HyperZ(Bolzonella et al., 2000)、Z-PEG(Le-Borgne et al., 2002)、IMPZ(Babbedge et al., 2004)、ZEBRA(Feldmann et al., 2006) 和 EAZY(Brammer et al., 2008) 等。模板匹配方法也称光谱能量分布 (Spectral distribution of energy，SED) 拟合方法，根据观测得到的光谱能量分布与红移之间的关系，建立红移预测模板。然后将观测得到的星等、颜色与模板进行对比，通常当 χ^2 值 (公式 1-4) 最小时，就认为得到了该天体的红移。其中所有使用的观测数据都需要进行消光改正。模板匹配方法受模板的限制，真实星系的模版大部分是较亮的低红星系，而星系合成的模板可能参数不完全正确。当然，经过近 20 年的积累，模板库也得到了很大改善。目前，模板匹配方法能够估计的最大红移达到了 6，预测的精度也有了明显提升。

$$\chi^2(z,T,A) = \sum_{i=1}^{N_{filt}} \left(\frac{F^i_{obs} - AF^i_{pred}(T,z)}{\sigma^i_{obj}}\right)^2 \tag{1-4}$$

式中，N_{filt} 表示观测的波段数，F^i_{obs} 表示在第 i 个波段的实测流量，$F^i_{pred}(T,z)$ 表示模板 T 中对应红移 z 处的流量，σ^i_{obj} 表示实测误差。

随着近年来大型光谱巡天望远镜的运行及机器学习技术的发展，机器学习方法开始在测光红移预测中的应用也越来越广泛 (Ball et al., 2007; Zhang et al., 2013; Way et al., 2006, 2009; Bonfield et al., 2010; Wang et al., 2007; Way et al., 2012; Kind et al., 2014; Carliles et al., 2010; Schindler et al., 2017)。红移估计属于机器学习的回归问题，根据已知天体的特征及红移值，通过学习算法进行训练，从而建立天体特征与红移之间的关系，并最终构建预测模型。通过预测模型，根据观测到的天体特征，就可以得到天体的红移。大量被准确证认的光谱红移增加了机器学习的训练样本，使得预测的精度比模版匹配方法更高，而且预测速度也较快。

红移估计方法的选择与已知样本的多少、是否具有代表性等问题密切相关。例如，对于低红移区域，机器学习方法明显要优于模板匹配，这是因为当具有足够多的具有代表性的训练样本时，构建的机器学习回归模型就会非常准确。但是对于覆盖了更大红移范围的深场巡天，模板匹配方法则更有优势 (Salvato et al., 2019)。因此，在实际应用中，要综合考虑数据特征和科学问题，再选择合适的红移测量方法，必要的时候，可以同时采取两种方法，优势互补。

1.3 选题意义和研究内容

1.3.1 选题意义

随着大型图像巡天（如 SDSS、PanSTARRS）、光谱巡天（如 SDSS、LAMOST）和其他波段巡天的开展，天文数据日益丰富和纷繁复杂。天文数据的独有特性为天文学家处理、分析和挖掘提出了前所未有的挑战。机器学习和人工智能已经成为数据科学时代的宠儿，为解决大数据问题提供了很好的解决途径。天文学家也不甘落后，将其应用在天文学领域的方方面面。天文研究重要的一环就是获得天体的光谱，通常光谱的获得代价昂贵。要想提高望远镜的观测和运行效率，精心挑选和计划输入星表是至关重要的。同时，研究天体的性质首先要确定天体的距离，对没有光谱的海量图像数据测得它们的测光红移具有重要的现实意义和科学应用价值。

大天区多目标光纤光谱天文望远镜（LAMOST；Large Sky Area Multi-Object Fibre Spectroscopy Telescope）是由我国自主设计的大视场光谱巡天望远镜。光学系统由一个 5.72 米 x4.4 米的反射施密特改正镜 MA，一个 6.67 米 x6.05 米的球面主镜 MB 及焦面三部分组成。由于在焦面上同时放置了 4,000 根光纤，理论上可以同时观测 4,000 个天体，获取它们的光谱，使得 LAMOST 成为当时世界上光谱获取率最高的望远镜。2021 年，LAMOST 开始启动二期工程，计划将光纤数量提升到 10,000 根，将重新成为世界上光谱获取率最大的望远镜。而且二期选址冷湖，具有最优秀的观测条件，将使 LAMOST 能够看得更远，为获取更大规模的类星体样本成为可能。LAMOST 任务之一是河外光谱巡天，为星系结构与演化的研究提供大量的信息，从而以更高的精度确定宇宙的组成和结构，加深对暗能量和暗物质本质的认识。

暗能量光谱巡天（Dark Energy Spectroscopic Instrument, DESI）在美国亚利

桑那基特峰国家天文台正式开始了第一期的巡天观测，为期五年。DESI 光谱巡天观测的首选目标是发射线星系、亮红星系及类星体。DESI 同样采用了大视场多光纤光谱采集技术，在焦面上配置了 5,000 根光纤，可以同时获取 5,000 个天体的光谱信息，成为目前世界上光谱获取率最高的望远镜。

为了提高望远镜的观测效率，优化输入星表是各个巡天项目的必备功课。结合图像和光谱巡天的需求，本文的研究目标主要包括三个方面：一是如何利用机器学习技术从大型图像巡天数据中选择高概率、高红移的类星体候选体；二是开展测光红移技术的研究，找到可靠的测光红移估计方法，为大型巡天设备的类星体和星系的测光红移计算服务；三是在线科研平台的开发和应用，结合天文应用需求，提高天文学家的工作效率。

1.3.2　研究内容

1.3.2.1　基于 BASS 巡天数据的类星体选源

BASS 和 MzLS 巡天的第三次释放数据含约 2 亿个数据源。将 BASS DR3 与光谱巡天 SDSS 和 LAMOST 数据交叉相关，获取已知样本的光谱类别。然后，将样本与 ALLWISE 数据库进行交叉匹配，获得已知样本的光学和红外信息，研究恒星、星系和类星体的多波段属性及其在多维空间的分布，也探讨各种不同类型的恒星的多波段属性及其在多维空间的分布。利用恒星、星系和类星体的已知样本，我们基于 XGBoost 算法构造不同的分类器：二元分类器和多元分类器。最后，应用由最优输入特征创建的优化分类器对 BASS DR3 星表中所有的源进行分类。各个源的分类标签和概率由不同的分类器给出。预选出来的类星体候选体可作为 LAMOST、DESI 或其他后续光谱巡天的输入星表。

1.3.2.2　基于大型巡天数据的类星体测光红移估测

基于 BASS 可见光巡天数据、ALLWISE 红外巡天数据、SDSS 和 LAMOST 光谱巡天数据，准备已知红移的类星体样本。对比 CatBoost、XGBoost 和随机森林在类星体测光红移估测时的性能以及在将样本分为低红移和高红移样本时的性能。提出红移预测的两步模型，即先将样本分成高红移和低红移两部分（以红移 3.5 为界），再分别对这两部分估测红移。比较一步模型和两步模型的性能。最后，我们将 CatBoost、XGBoost 和随机森林中性能最为优越的算法作为分类和红移估测的核心算法，设计一个预测类星体测光红移的流程，用于预测 BASS 巡天数据中所有类星体候选体的红移。

1.3.2.3　基于大型巡天数据的星系测光红移估测

基于 DESI 可见光巡天数据、WISE 红外巡天数据、SDSS 和 LAMOST 光谱巡天数据，研究星系的测光红移估测方法。应用模板匹配（EAZY）和机器学习（CatBoost）两种方法进行预测。对比这两种方法在星系红移预测的性能及适用条件。机器学习方法受限于已知样本，超出红移范围的红移预测以模板匹配预测

为准。最后用 CatBoost 构建的最优模型和 EAZY 对全部 DESI DR9 数据的星系进行测光红移预测。

1.3.2.4 面向类星体选源与红移估计的在线计算平台

大规模的数据对于类星体选源与红移估计的计算能力提出了巨大的挑战。我们基于虚拟天文台的接口及协议框架，开展了面向类星体选源与红移估计的在线计算平台技术研究。在线计算平台将提供按需定制的在线科研环境，利用计算与数据的融合，避免了大规模数据的下载、传输。同时，在数据准备、模型应用等不同环节进行计算的并行化设计，从而提升了大规模数据的处理速度。

1.3.3 论文结构

本文第一章介绍了类星体的发现和类星体的预选源方法，以及测光红移和测光红移技术；第二章详细阐述天文巡天及数据情况及机器学习技术在大数据时代的研究应用现状；第三章着重介绍基于 BASS 巡天数据的类星体选源工作；第四章着重介绍类星体的测光红移预测，高红移类星体可以揭示宇宙早期的信息，始终是天文学家追逐的热点；第五章着重介绍星系的测光红移预测；第六章主要描述在科学平台方面的一些研究工作，以应对科研工作中对海量数据的高效分析与处理问题，是大数据时代天文学研究必不可少的关键技术之一。随着望远镜技术、信息技术及其相关技术的不断发展进步，天文学观测进入到了巡天时代，天文数据以前所未有的速度增长，传统的科研模式将发生质的变化。

第 2 章　天文大数据与机器学习

宇宙，遥远而神秘。从古至今，人类对探索宇宙、了解宇宙有着不懈的追求。在远古时期，受限于条件的限制，人们只能通过双眼观看星空，用原始的笔和纸来记录，从而发现天体的秘密。1609 年，意大利天文学家、物理学家伽利略发明了人类历史上第一台天文望远镜，开启了利用望远镜进行天文观测的新时代，使得人类对于宇宙的认识有了巨大的进步。然而，宇宙是如此浩大，历经千年探索，人类对宇宙的了解仍然只是冰山一角，天文学家期待看得更远、更清晰、信息更丰富的望远镜的出现。

2.1　天文观测进入大型巡天时代

20 世纪中后期以来，各种技术正在经历史无前例的飞速发展，包括探测器和空间技术、望远镜制造技术、大面积探测阵列技术、计算机信息技术、网络技术等，使得大型天文望远镜的建设进入快速发展时期。预计到 2025 年，全球口径大于 6 米的光学望远镜将有近 30 台 (Cai et al., 2021)。此外，还有许多射电及空间望远镜正在或即将投入运行，天文观测进入了一个全波段的数字巡天时代。

光学巡天：

斯隆数字巡天 (Sloan Digital Sky Survey, SDSS)[1] 是目前世界上最为成功和最有影响力的天文巡天项目之一。项目的前两期巡天主要使用位于美国新墨西哥州 APO 天文台 (Apache Point Observatory) 建造的一台口径为 2.5 米的专用天文望远镜。此望远镜配备了两台强大的天文专用仪器: 一是配备了专用的 CCD 相机，可以一次拍摄 1.5 平方度的天空；二是配备了光纤光谱仪，可以一次拍摄超过 600 个天体目标。SDSS 巡天开始的几周所获得的数据就超过了过去所有望远镜观测所积累的数据总和。从第三期巡天开始，同时使用了智利安第斯山脉的拉斯坎帕纳斯天文台的 2.5 米艾琳妮杜邦望远镜。截止目前，SDSS 完成了 4 期巡天，共发布了 17 版数据 (Stoughton et al., 2002; Abazajian et al., 2003, 2004, 2005; Adelman-McCarthy et al., 2006, 2007, 2008; Abazajian et al., 2009; Aihara et al., 2011; Ahn et al., 2012, 2014; Alam et al., 2015; Albareti et al., 2017; Abolfathi et al., 2018; Aguado et al., 2019; Brad et al., 2020; Abdurro'uf et al., 2022)。最新发布的 DR17，数据总量达到 652TB，包含了五亿个恒星和星系的测光数据，以及接近 4 百万条的光谱数据。

大天区多目标光纤光谱巡天望远镜（Large Sky Area Multi-Object Fiber Spectroscopic Telescope, LAMOST），也称郭守敬望远镜 [2]，是一架由我国自主研制的大视场兼备大口径 (有效口径 3.6-4.9 米) 巡天望远镜 (Cui et al., 2012)。它采用独

[1] https://www.sdss.org

[2] https://www.lamost.org

创设计，使得一次观测可以同时获取 4,000 个天体的光谱，具有极高的光谱获取率，被称为“光谱之王”。2015 年发布的第一版数据 (Luo et al., 2015)，就包括了近 200 万条光谱，截止到 2021 年 3 月，LAMOST 共发布了超过一千多万条光谱数据，是世界上最大的天体光谱数据库，在恒星、银河系等领域取得了许多重大成果。

暗能量光谱巡天 (The Dark Energy Spectroscopic Instrument, DESI) 的图像巡天[3] 主要是为光谱巡天提供侯选天体。DESI 图像巡天包括三个子巡天：一个是位于智利 CTIO 的 DECam 巡天；另两个则是北京-亚利桑那（BASS）巡天和 MzLS 巡天，总覆盖天区超过 14,000 平方度。2019 年，DESI 图像巡天发布了第九批测光数据 (Dey et al., 2019), 其中图像数据超过 100TB，星表包括了近 20 亿天体测光信息。

中国巡天空间望远镜（The China Space Station Telescope, CSST）是正在建设中的一个空间光学望远镜，望远镜口径 2 米，观测视场超过 1.1 平方度，角分辨率小于 0.13 角秒，覆盖波段包括近紫外、u、g、r、i、z、y 及短红外波段，计划 2024 年开始投放运行。预计 10 年巡天将观测 17,500 平方度的天区，获取数十亿天体近 5PB 的原始数据。主要科学目标包括暗能量、暗物质和宇宙结构形成和演化；星系起源与演化；活动星系核和超大质量黑洞的形成和演化；太阳系外行星、天体测量和太阳系天体研究。

大型综合巡天望远镜 (The Large Synoptic Sky Survey Telescope, LSST)[4] 是已经正在建设中的下一代地基全天望远镜 (Tyson, 2002; LSST Science Collaborations, 2009)。它将座落在智利的塞罗 · 帕拉纳 (CerroParanal) 观测站，该地被认为是世界上最好的天文观测地点之一。LSST 采用的是 8.4 米口径的主镜，有效口径是 6.7 米，计划产生 6 个波段的数据 (0.3-1.1 微米), 深度覆盖南天 20,000 平方度的天区。与以往巡天项目不同的是 LSST 拍摄能力非常强劲，配备的探测器阵列 (3.2 Gigapixel) 可以每次曝光 9.6 平方度，且每个像素分辨率达到 0.2 角秒，每晚产生超过 800 张的全景图像，这样它可以在一周内对全部观测天区扫描 2 次，让天文学家可以对观测目标进行时域研究。LSST 预计每晚产生 30TB 的数据，而整个巡天计划将会产生 60PB 的数据以及超过 200 亿行的巨大星表。

红外巡天：

两微米全天巡天（The Two Micron All-Sky Survey, 2MASS)[5] 是一个近红外波段的全天巡天项目，包括 J、H 及 K 三个波段。2MASS 配备了两台高度自动化的 1.3 米望远镜，能够同时在 J（1.25μm)、H（1.65μm）及 K（2.17μm）3 个波段观测天空。数据已于 2002 年全部释放。2MASS 星表包含了近 3 亿颗恒星、50 万星系及星云的天体测量及测光数据及超过 12TB 的图像数据。

广域红外巡天望远镜 (The Wide-field Infrared Survey Explorer, WISE)[6] 是由

[3]https://www.legacysurvey.org

[4]https://www.lsst.org

[5]https://www.ipac.caltech.edu/project/2mass

[6]https://www.nasa.gov/mission_pages/WISE/main/index.html

NASA 资助的空间探测器 (Wright et al., 2010)，它于 2009 年启动，目标是获取全天在 3.4μm、4.6μm、12μm 和 22μm 四个波段的信息。四个波段的角分辨率分别达到 6.1、6.4、6.5 和 12.0 角秒。2012 年，在已有 WISE 图像数据的基础上，ALLWISE 星表数据库正式发布，包含了 7 亿多个天体的中红外数据。2019 年，结合 WISE 及 NEOWISE 的所有的全天观测图像，基于全新的数据分析软件 CATWISE，发布了新一版的星表数据 (Marocco et al., 2021)。此时已包括了近 19 亿天体红外测光数据。中红外波段数据在类星体识别上扮演重要角色，随着新一代 6.5 米口径的红外空间望远镜詹姆斯 · 韦布（The James Webb Space Telescope，JWST）的运行，将提供更大规模天体的红外数据。

射电巡天：

500 米口径球面射电望远镜 (The Five-hundred-meter Aperture Spherical radio Telescope, FAST)[7] 是中国建设的目前世界上第一大的单口径射电望远镜 (Jiang et al., 2019)，其拥有相当于 30 个标准足球场大的接收面积，覆盖了从 70MHz 到 3GHz 的可观测频段。FAST 配备了 7 套数据接收机及脉冲星、谱线、SETI 等多种数据记录终端，每秒最高采集的数据将可达到 38GB，每年的数据量为 20PB。

平方公里射电阵（The Square Kilometer Array, SKA）是一个国际合作在建的射电天文项目，将成为世界上最大的射电望远镜。中国于 2021 年正式批准《成立平方公里阵列天文台公约》，从而成为 SKA 的正式成员。SKA 是由部署在上千公里跨度范围内的数千个碟形天线构成的综合孔径阵列，集大视场、高灵敏度、高分辨率、宽频率范围等于一身，其科学研究目标包括第一代天体如何形成、星系形成与演化、暗能量性质、宇宙磁场、引力本质、生命分子与地外文明等许多天文前沿科学问题。然而，SKA 所面临的数据处理挑战超乎想象，每秒的观测数据量将达到 PB 量级。SKA 数据的深度分析和加工将分布于几大洲的区域数据中心完成。按照 SKA 的数据流规模，估计在建设的 SKA1 需要输送到区域数据中心进行深度分析的科学数据就达到了每年 300PB。到了 SKA2 阶段，从 SKA 天文台产生的预处理数据的规模将扩展到 SKA 先导项目的 100 倍以上，达到 EB 量级。因此，SKA 将给数据存储、数据处理、数据可视化、数据分析等各个阶段都带来了前所未有的挑战。

此外，还有包括三十米望远镜（Thirty Meter Telescope，TMT）、欧几里得空间望远镜（The Euclid Space Telescope）、大视场巡天望远镜（The Wide Field Survey Telescope，WFST）、司天计划等多个正在建设中的望远镜项目。随着这些大型望远镜的投入运行，天文观测从可见光、射电波段扩展到包括红外、紫外、X 射线在内的电磁波各个波段，形成了全波段天文学。表 2-1 列出了部分望远镜产生的数据规模。

天文全波段的巡天观测导致天文科学数据急剧增加。海量天文数据为天文学家带来了有关天体的丰富信息，对于天文学家作出创新性的发现提供了新的机遇。然而，大型巡天项目也正使得天文学变成了一门数据密集型、计算复杂型

[7] https://fast.bao.ac.cn/

表 2-1 多个望远镜项目的数据规模。

Table 2-1 Data scale of multiple telescope projects.

巡天项目	数据规模	天体数量规模
2MASS	10TB	5 亿
GALEX	30TB	3 亿
WISE	10TB	7 亿
SDSS （DR16）	100TB	12 亿
GAIA (DR3）	150TB	18 亿
DESI （DR9）	110TB	20 亿
CSST	预期 5PB	100 亿
LSST	预期 150PB	200 亿

的学科。面对海量的天文科学数据，需要有新的数据处理算法以及支持大规模数据存储、计算的科研环境。

2.2 数据挖掘技术及其在天文学上的应用

20 世纪 90 年代，随着多样化的信息系统的建设普及，数据库技术进入到一个全新阶段。数据库从管理一个简单数据发展到管理各种计算机所产生的图形、图像、音频、视频、电子档案、WEB 页面等多种类型的复杂数据，数据量也越来越大。人们迫切希望能从这些收集到的数据中发现隐藏的信息，从而催生了数据挖掘技术的出现。比如早期的决策支持系统、专家系统等都是数据挖掘技术应用的形式。简单而言，数据挖掘（Data Mining，DM）(Han et al., 2006; Witten et al., 2011) 是人们在处理数据、分析数据和研究数据所形成的一套完整的方法与工具，涉及数据库和数据仓库技术、统计分析、机器学习、模式识别、神经网络、数据可视化等等。数据挖掘技术可以有效地帮助我们从大数据中发现规律，形成知识。

数据挖掘技术的发展伴随着数据的不断增长与应用需求的不断递增，是一个逐渐演变的过程。数据挖掘的出现就是源于人们试图实现基于数据的自动决策，因此机器学习在一开始就成为了人们关注的焦点。机器学习的过程是将一些已知信息作为范例输入计算机，机器通过学习这些范例并生成相应的规则，从而解决同类问题。但由于已知数据还不充分，在很多应用领域并没有达到理想的效果。同时，随着神经网络技术的形成与发展，知识工程技术受到人们的关注。知识工程不同于机器学习，而是直接给计算机输入已被代码化的规则，让计算机通过使用这些规则来解决某些问题，专家系统就是这种方法的产物。后来，由于数据的不断积累与机器学习技术的发展，许多传统的机器学习算法在实际应用中取得了很好的效果，数据挖掘技术的核心重新回到机器学习上来。

目前，数据挖掘技术在很多领域都得到了非常广泛的应用，通过总结这些行业应用，可以发现数据挖掘技术主要侧重解决的问题包括关联分析 (Association analysis)、聚类分析（Clustering）、分类（Classification）、回归（Regression）、时序分析（Time-Series analysis）等。

(1) 关联分析（Association analysis）：若两个或多个变量的取值之间存在某种规律性，则称为关联。关联分析的目的是找出数据中隐藏的关系，从而揭示数据背后的内在规律。关联分析，即利用关联规则进行数据挖掘。关联分析在商业领域应用广泛，比如通过对顾客的购买行为进行分析，可以发现 “90% 的顾客在一次购买活动若购买了 A，则也会购买 B” 之类的知识。关联分析可分为简单关联、时序关联、因果关联等。

(2) 聚类分析（Clustering）：聚类分析是指通过对数据记录的分析，寻找数据之间的规律，从而建立一定的分类规则来确定每个数据记录的类别。分类规则的建立是通过聚类分析的工具及算法所决定的。俗话说，“物以聚类”，同一类数据具有一定的相似性，聚类分析就要从进行分析的数据中找到这种相似性。

(3) 分类分析（Classification）：在选择的已知数据中包括了不同类别的数据，这个已知数据集也称为训练集，每个数据具有一个所属类别的标签。分类分析就是通过对训练集中的数据进行分析，建立分类模型，然后利用这个分类模型对未知分类标签的数据进行分类。要构建分类器，需要有一个训练样本作为输入，训练样本越丰富，分类的准确率也就越高。因此，在当前的大数据时代，分类在许多商业与科学领域都有着广泛的应用。

(4) 回归分析（Regression analysis）：回归分析指的是确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法。在数据挖掘应用中，回归分析算法基于目标及已知数据样本建立预测模型。与分类分析所不同的是，回归的目标不是离散的类型标签，而是一组连续数值。

(5) 时序分析（Time-Series analysis）：在现代社会的生产生活中，存在着大量的时间序列数据，如股票价格、各种汇率、销售数量、产品的生产能力、天气数据等，这些数据记录了各个时刻的重要信息。时序分析就是通过对这些数据的分析，发现各时间序列数据之间的相互关系，从而提高人们对这类系统的认识和理解，并最终对未来时刻可能出现的情况进行有效地预测和控制。

天文学是一个数据科学，从大规模的观测数据中寻找新的天文知识，验证天文理论是天文研究的主要方式。自 20 世纪 90 年代起，天文学家便开始了在天文科研中应用机器学习方法的探索，在 2004 年逐步形成规模。尤其是近年来，天文数据的急剧增长，数据挖掘等大数据分析方法与工具成为了天文学研究的重要部分。表 2-2 列出了天文科研中应用数据挖掘算法来解决的一些主要问题。

从表 2-2 可知，天文中的数据挖掘任务主要体现在四个方面，包括分类、回归、聚类及时序分析。对应每一类任务，都有很多可以采用的学习算法，每一种算法都具有各自适用的场景。同时，由于数据规模与计算能力的不断增长，传统算法也在不断发展与演变。下一节将介绍当前主流的学习算法。

表 2-2 利用数据挖掘算法解决的主要天文问题及具体算法。

Table 2-2 Main Astronomical Problems Solved by Data Mining Algorithms.

数据挖掘任务	天文应用	主要方法
分类	光谱分类	人工神经网络、支持向量机、学习向量化
	图像分类	K 近邻、决策树、旋转森林、极限学习机
	星系形态分类	贝叶斯网络、随机森林、极端学习树
	多波段数据分类	C4.5、CART、XGBoost、CatBoost、
	太阳活动	逻辑回归、深度学习
回归	星系测光红移	人工神经网络、支持向量回归、核回归、决策树
	类星体测光红移	主成分回归、高斯过程、K 近邻回归、深度学习
	恒星物理参数估计	小二乘回归、决策树、随机森森、XGBoost、CatBoost
聚类	天体分类	主成分分析、DBSCAN、K-Means、OPTICS、Cobweb
	稀有目标搜寻	高斯混合建模、自组织映射、t-SNE
	新天体搜寻	凝聚层次聚类、期望最大化、深度聚类
时序分析	趋势分析	随机森林、XGBoost、LightGBM、支持向量机、深度学习
	新天体探测	人工神经网络、周期因子、自回归模型 AR
	变源分类	移动平均模型 MA、自回归移动平均模型 ARMA

2.3 机器学习算法

随着数据规模的不断增长，机器学习算法得到了快速的发展和应用，成为数据挖掘任务的首选技术。机器学习的主要目标是研究如何让计算机模拟人类的学习行为，通过经验自动提高算法效率，从数据中学习隐含的模式并建立模型，从而能够对相似的问题做出预测 (Mitchell, 1997)。通常，机器学习根据模型的不同分成四类：监督学习、非监督学习、半监督学习和强化学习。监督的机器学习算法将包含有标签的数据作为训练集进行模型构建。训练集中的每一个数据实例包含一组特征和目标标签，训练样本数据越丰富、越全面，构建的模型就越精确。非监督学习算法则是训练样本没有给出标签，直接从数据集中寻找数据之间的关系。监督的学习算法主要应用于分类与回归问题中，比如支持向量机、随机森林、XGBoost、CatBoost 等都属于监督学习算法，而非监督的学习算法通常应用于聚类、降维及离群检测等相关问题中，如表 2-2 聚类问题中所列算法都属于非监督的学习算法。半监督学习模型中的训练样本大部分是没有标签的，只有很小部分有标签。在实际应用中，我们拥有的数据量足够多，而有标签的数据却相当少，如果为数据都给出标签是很耗费时间和精力的。因此与监督学习模型相比，半监督学习模型在缺乏标签数据的情况下较常被采纳，准确度还是比较高的。强化学习则是利用奖惩函数让未标记的数据极可能地接近真实的目标。在本研究工作中，我们主要利用了监督的学习方法，下面对一些主流的监督学习算法作个简单介绍。

2.3.1 支持向量机

支持向量机（Support Vector Machine，SVM）(Vapnik, 1995) 是最流行的监督学习算法之一，其基本思想是在高维空间内利用线性函数的对偶核，并通过内积

空间的向量运算来处理线性不可分的数据。给定一个具有 N 个特征的训练数据集，支持向量机将在 N 维空间里寻找一个超平面，使得不同类别的数据能够很好地被分开。例如：在特定的二维空间中，这个超平面就是一条直线，把整个平面分成了两部分，直线两边分别代表一个分类。一般情况而言，最优超平面是指所有数据点到此平面的距离最远，那么这个超平面就成了一个决策边界，可以用于分类预测。

但是在很多情况下，多维空间的数据并不是线性可分的。支持向量机引入了核函数的方法，将输入数据集转换到特征空间，而在转换后的特征空间中，不同类别的数据就变成线性可分了，然后再来寻找最优超平面。核函数有多种形式，比如多项式映射、高斯径向基函数、指数径向基函数、傅立叶级数、样条函数、B 样条函数、叠加的核函数、张量积等。在实际应用中需要根据数据及分类情况进行选择。SVM 算法主要应用于天体分类中，包括星系分类 (Huertas-Company et al., 2008)、星系与恒星分类 (Fadely et al., 2012; Krakowski et al., 2016)、恒星与 AGNs 分类 (Peng et al., 2012; Malek et al., 2013)、恒星分类 (Ksoll et al., 2018) 等。但由于其较慢的训练速度，不能满足大数据时代的需求。

2.3.2　决策树

决策树是一个通过训练构建的非参数化模型，可以采用一颗自上而下的树形图来描述，故称为决策树。决策树可以用于分类与回归任务。决策树由一系列的决策点组成，每个决策点表示一个条件，最终的叶子结点则表示一个分类。决策树的训练过程就是树的生成过程。训练过程就是根据一定的标准和规则分割训练样本集为几个子集，然后再以相同的规则去分割每个子集，递归这个过程，直到每个子集只含有属于同一类的样本时停止。分割的标准称为模型超参数，主要包括基尼不纯度和信息增益。决策树的主要优点是易解释，分类速度快，但同时也具有不稳定性，易于过拟合等不足。由于决策树的显著优点及缺点，人们就在思考如何发挥它的优点而减少缺点。集成方法为解决这个问题提供了一种创造性的思路。集成方法是指联合几个弱监督学习方法到一个预测模型中，从而实现预测性能出现显著改善的新的强学习方法。集成方法有两种形式：一种是 “Bagging”，随机森林是典型代表，基本的过程如图 2-1 所示; 另一种是 “Boosting” 模式，典型代表包括梯度决策树、LightGBM、XGBoost、CatBoost 等，基本原理如图 2-2 所示。集成方法中的弱学习器可以是决策树，也可以是其它学习方法。

2.3.3　随机森林

随机森林 (Random Forests, RF)(Breiman, 2001) 是一种由决策树集合组成的集成学习方法。随机森林从训练集中随机选择若干个子样本，然后分别对每一个子样本进行决策树的构建。每棵树训练使用的特征也是随机选择，采用相同的标准进行训练。随机森林中决策树的总数及树的最大深度通过模型超参数来进行设置。随机森林总的性能由每棵树求平均而得，从而避免了像单棵决策树出

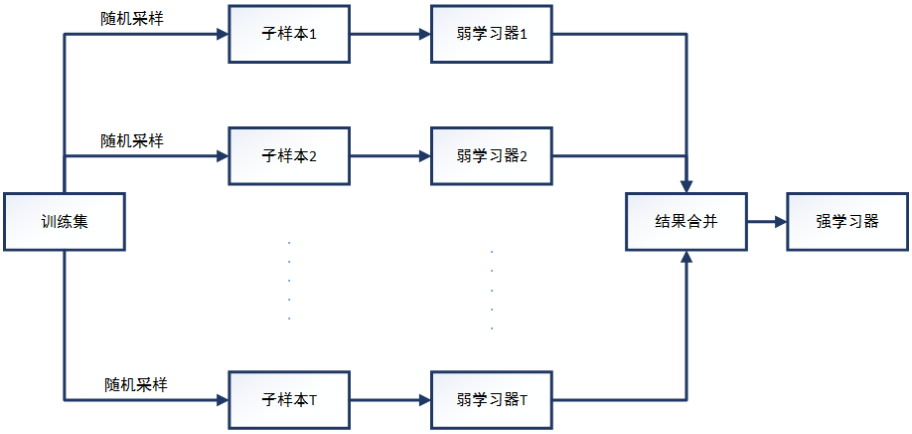


图 2-1 基于 Bagging 的集成学习方法原理。

Figure 2-1 The principle of bagging ensemble method.

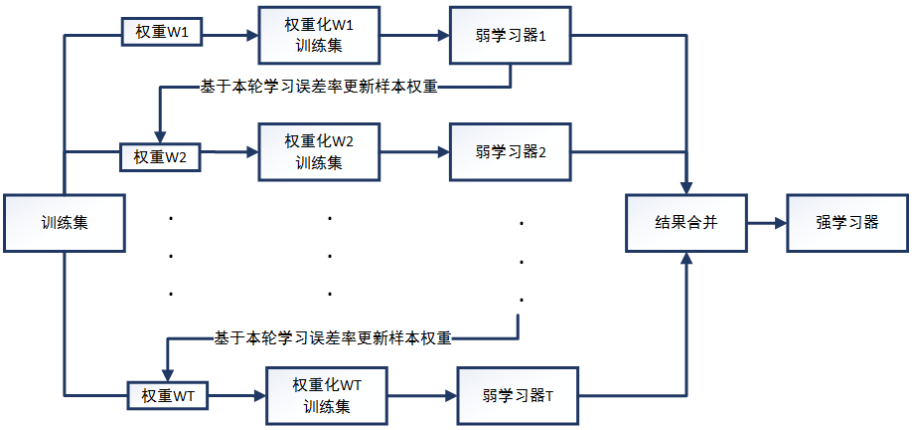


图 2-2 基于 Boosting 的集成学习方法原理。

Figure 2-2 The principle of boosting ensemble method.

现过拟合现象。随机森林即可作为监督的学习方法用于分类与回归任务，也可用作非监督的学习方法进行奇异数据检测。随机森林在天文领域具有广泛的应用，例如：Arnason 等 (Arnason et al., 2020) 利用 RF 分类算法搜寻 X 射线源双星；Pichara 等 (Pichara et al., 2012) 利用 RF 分类算法搜寻类星体；Carliles 等 (Carliles et al., 2010) 利用 RF 回归算法进行测光红移估计；Plewa(Plewa, 2018) 用于恒星分类；Ishida 等 (Ishida et al., 2019) 用于超新星分类等。

2.3.4　梯度提升决策树

梯度提升决策树（Gradient Boosting Decision Tree，GBDT）(Jerome, 2001) 是一种基于“Boosting”模式的集成学习方法，采用决策树作为基学习器，每一次迭代训练是在之前训练的基础上实现更好的拟合。模型结果为一系列 CART 树的集合：T_1，...，T_n。最终预测结果为每棵树的结果之和，公式表示如下：

$$\bar{y} = \sum_{n=1}^{N} f_n(x), \quad f_n \subset \Gamma \tag{2-1}$$

$\bar{y}$ 代表预测结果，f_n 表示第 n 次迭代生成树的预测函数。

2.3.5　XGBoost

XGBoost （eXtreme Gradient Boosting）(Chen et al., 2016) 是一种开源的梯度提升算法，于 2016 年正式发布。XGBoost 具有良好的性能，得到了机器学习社区的认可。它是由基于梯度提升决策树（Gradient Boosting Decision Tree，GBDT) (Jerome, 2001) 发展而来，属于集成学习算法，通过学习多个弱分类器（决策树），经过不断迭代而构建强分类器。XGBoost 即可用于分类问题，也可解决回归问题。与 GBDT 相比，在计算残差时，XGBoost 会综合考虑一阶和二阶导数，而 GBDT 只使用一阶导数，这是两者的明显区别。XGBoost 的目标函数可以写成如下泰勒函数展开式：

$$Obj^{(t)} = \sum_{i=1}^{N} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \tag{2-2}$$

式中，g_i 是损失函数的一阶导数，h_i 是损失函数的二阶导数。Ω 惩罚项，定义了树的复杂度，具体可表示为：

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{k=1}^{T} \omega_k^2 \tag{2-3}$$

式中，T 表示叶子的数量，ω_k 表示第 jth 片叶子的权重。公式 (2-3) 有助于避免过拟合。

对于一个具有 n 个样本，m 维特征的数据集 $\mathcal{D} = \{(x_i, y_i)\}(|\mathcal{D}| = n, x_i \in \mathbb{R}^m, y_i \in \mathbb{R})$。假定最终模型有 k 颗树组成，对样本 x_i 的预测按如下公式计算：

$$\hat{y_i} = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathbb{F} \tag{2-4}$$

式中，f_k 代表一颗决策树，$f_k(x_i)$ 表示 x_i 在第 k 棵树上获得的分数，$\mathbb{F}$ 表示所有可能的决策树。

XGBoost 算法的有效性已在许多机器学习和数据挖掘挑战中得到广泛认可。Bethapudi 等 (Bethapudi et al., 2018) 采用 XGBoost 算法从脉冲中区分出脉冲星信号；Mirabal 等 (Mirabal et al., 2016) 用于对 FermiLAT 星表中未知源的分类；金鑫等 (Jin et al., 2019) 从 PanSTARRS1 数据集中进行类星体选源；付煜铭等 (Fu et al., 2021) 将其用在搜寻银道面类星体候选体研究工作中。

2.3.6 CatBoost

CatBoost(Dorogush et al., 2018) 也是由 GBDT 发展而来的梯度提升集成学习方法，由俄罗斯的搜索引擎公司 Yandex 开发，是一个高性能的开源学习算法，同时支持分类与回归任务。与其它 GBDT 算法不同的是，CatBoost 采用了对称决策树作为基学习器，对称决策树是一棵完全二叉树，所有非叶子节点具有相同的分类标准，这种设计有助于加快训练速度以及避免过拟合。CatBoost 直接支持类别特征，而不需要进行预处理。CatBoost 采用有序的目标统计（Ordered Target Statistics）方法将类别特征转换为数值型特征。

此外，CatBoost 在采用缺省参数的情况下仍然可以获得较好的结果，从而节省参数优化时间。CatBoost 也支持 GPU 计算，可以进一步加速对大规模样本数据的处理。CatBoost 一经推出，受到业界的极力推崇，在金融、保险、医疗及石油等领域都具有广泛的应用。在天文领域，向关杰等人利用 CatBoost 来进行恒星参数的估计 (Xiang et al., 2021)；Coronado-Blâzquez (Coronado-Blâzquez, 2022) 应用 CatBoost 对未认证的伽马射线源进行了分类, 获得了很好的效果。作为一种较新的集成算法，具有很大的应用潜力。

2.3.7 人工神经网络

人工神经网络（Artificial Neural Network，ANN）是一组结构模糊的算法，其灵感来源于构成人脑的生物神经网络。人工神经网络所具有的灵活结构和非线性使其能够执行多种任务，包括分类、回归、聚类等。人工神经网络通常由输入层、输出层及多个隐藏层组成。图 2-3 展示了一个神经网络架构。由图可知，此网络由输入层、输出层及二个隐藏层组成，输入数据从输入层，通过隐藏层，到达输出层。相邻层之间的各神经元实现全连接，但同层各神经元之间无连接。网络中每一个神经元的值（输入层除外）是前一层神经元的线性组合, 在此基础上再应用一个非线性的激活函数。隐藏层的数量及神经元的数量通过超参数设置，而输入层及输出层的神经元的数量根据学习任务（分类或回归）来定义。多层感知机（Multilayer Perceptron, MLP）是最简单的人工神经网络算法实现。人工神经网络在星系形态分类 (Storrie-Lombardi et al., 1992)、光谱分类 (Singh et al., 1998; Snider et al., 2001) 、测光红移估计 (Firth et al., 2003; Tagliaferri et al., 2003; Vanzella et al., 2004)、星系参数估计 (Teimoorinia et al., 2016)、光谱参数估计 (Das et al., 2019) 等都有成功案例。近年来，随着人工神经网络向深度学习发展，传统

神经网络算法的应用也在逐渐减少。

2.3.8　深度学习

深度学习（Deep Learning）是由人工神经网络发展而来的新型学习算法。它与传统机器学习的区别主要在于特征表示建立的过程。传统机器学习算法通常需要根据专家知识来设计及优化特征，而深度学习通过多层神经网络的架构从原始数据中逐层提取抽象特征。因此，深度学习更倾向于从原始数据学习而无需太多数据的后期加工。深度学习具有强大的特征学习能力，尤其在计算机视觉和语音识别领域。主流的深度学习框架包括 TensorFlow、Keras、PyTorch，基于这些框架，很容易构建和训练深度神经网络预测模型。当前主流的深度学习网络模型有卷积神经网络模型（Convolutional Neural Network，CNN）、循环神经网络（Recurrent Neural Network，RNN）、生成对抗网络（Generative Adversarial Networks，GANs）、强化学习（Reinforcement Learning，RL）等。深度学习很早就开始应用于天文领域，比如图像与光谱分类 (Hâla, 2014)、恒星参数测量 (Parks et al., 2018; Fabbro et al., 2018)、搜寻系外行星 (Pearson et al., 2018; Shallue, 2018) 及测光红移估计 (Pasquet-Itam et al., 2018) 等。

2.4　本章小结

随着天文观测进入大型巡天时代，天文数据继续不断增长，机器学习技术在天文学的诸多领域越来越发挥着重要的作用。而机器学习算法本身也在经历从简单的单个算法到集成算法，再到深度学习的不断进化过程中，以适应不同的问题场景，优化预测的性能及尽可能短的建模时间。可以认为，集成算法的提出与实现是机器学习发展的一个里程碑，极大地推动了传统机器学习算法的提升。在Henghes et al. (2021) 的工作中，他们基于相同的样本对比了逻辑回归、最近邻、决策树、梯度决策树、随机森林、多层感知机等多种算法在星系测光红移预测方面的性能，综合考虑预测性能和训练时间，结果显示梯度决策树性能最优，这也体现了集成学习算法的优越性。虽然支持向量机、最近邻及决策树等古老的算法通过不断改进，优化以适应大数据训练的要求，如增加并行计算的能力，实现 GPU 算法等，但在诸多同类问题的对比中，集成学习的 XGBoost、CatBoost 方法要比传统方法更为优越，并行化的能力也更强，这将在本文的第三 ~ 第五章的实际应用中进一步得以验证。对于目前被热捧的深度学习算法，它的优越性主要体现在自动化的特征分析上，深度学习方法通常需要有较大的算力支持。因此，深度学习主要应用在直接对原始观测图像进行目标学习的问题中，如果在输入特征固定且特征数量不多的情况下，深度学习方法并不能很好地发挥出它的优势。因此，本研究主要采用 XGBoost、CatBoost 方法进行建模，同时也采用其它传统方法进行对比验证。

第 3 章　类星体选源

目前，类星体证认的主要方法是拍摄它们的光谱，通过谱线特征，最终确认是否是类星体。然而，光谱拍摄的成本很高。截止目前，我们总共获取到的光谱数据还不到 2000 万条。相对于光谱，天体的测光信息却较容易获得，正如在第二章所述，大型巡天设备观测到不同波段的大量测光数据。例如，PanSTARRS、CatWISE(Marocco et al., 2021)、DESI 图像巡天数据 DR9(Dey et al., 2019)、GAIA DR3 等多个数据集都已将近 20 亿。因此，类星体选源问题就是如何从这些大规模的测光数据中，高效地挑选出具有大概率是类星体的侯选体，然后再将这些侯选体提供给光谱望远镜进行观测，从而使得在相同的光谱望远镜时间下，尽可能地发现更多的类星体，这就需要有高效准确的分类模型。本研究工作主要基于 BASS DR3 的测光星表数据进行分类模型训练，从而构建最优的分类模型，并利用这些模型从 BASS DR3 的星表中甄选高置信度的类星体候选体。

3.1　数据

北京—亚利桑那巡天项目（BeiJing-Arizona Sky Survey，BASS）(Zou et al., 2017a) 是一个由中科院国家天文台与美国亚历桑那大学共同负责的多色图像巡天项目，是 DESI 图像巡天的重要组成部分。BASS 巡天使用位于 Kitt Peak 的一架口径 2.3 米的望远镜，覆盖了北银纬 30 度以北约 5,400 平方度的天区。BASS 巡天自 2015 年 1 月开始，到 2019 年 3 月结束，共观测了 250 个夜晚，主要观测波段为 g 和 r。同时，另一个巡天计划 MzLS(The Mosaic z-band Legacy Survey), 也是 DESI 图像巡天的一部分，与 BASS 巡天天区基本一致。MzLS 采用的是口径 4 米的望远镜，主要波段为 z，MzLS 巡天从 2016 年 2 月开始观测，到 2018 年 2 月结束，共 383 个夜晚。BASS 巡天共进行了三次数据发布，第一版数据 (Data Release 1, DR1)(Zou et al., 2017b) 于 2017 年 6 月发布，只包括了 g、r 两个波段的图像及测光信息。2018 年，第二版（DR2）(Zou et al., 2018) 数据发布，第二版数据中包括了 MzLS 巡天的 z 波段的数据。然后在 2019 年，又发布了第三版数据（DR3)(Zou et al., 2019)，DR3 中包括了这两个计划中的所有数据，其中星表包括单次测光及叠加测光星表，叠加测光星表具有更好的深度，g、r、z 三个波段的极限星等分别达到了 24.2、23.6、23（AB 星等）。考虑到数据的可靠性，我们只对极限星等范围内的数据进行了预测。

SDSS 是目前为止最为成功的天文巡天项目之一，2020 年进行了第 16 版的数据发布（DR16）(Ahumada et al., 2020), 也是第四阶段巡天观测的第一次数据发布，包括光谱和测光信息。同年，Brad 等 (Brad et al., 2020) 对 DR16 的类星体进行了再次确认，单独发布了类星体星表 DR16Q，包括了 SDSS 四个阶段发现的所有类星体，其中有 225,082 个类星体是首次发布。我们的已知样本主要来自

表 3-1 所用研究数据的下载地址。

Table 3-1 Websites for catalogues.

BASS-DR3 catalogue https://nadc.china-vo.org/data/data/bassdr3coadd/f
Known stars, galaxies and quasars from SDSS http://skyserver.sdss.org/dr16/en/tools/search/sql.aspx
Known stars, galaxies and quasars from LAMOST http://dr5.lamost.org/v3/catalogue
SDSS DR16 Quasar catalog (DR16Q) https://www.sdss.org/dr16/algorithms/qso_catalog

DR16 的光谱星表 SpecObj 及 DR16Q。为了保障数据的可靠性，选择的天体数据要求 zWarning=0，共得到 880,652 个恒星、2,616,381 个星系和 749,775 个类星体。

LAMOST 作为世界上最大的光谱数据库，至今已获取了 1 千多万条光谱数据。我们选择了于 2017 年发布的第五版数据（DR5），共包括了 152,863 个星系、52,453 个类星体和信噪比 S/N 大于 10 的恒星 7,146,482 颗。

此外，由于类星体在红外波段有较明显的特征，我们从 ALLWISE 数据集中选择中红外波段的测光数据。ALLWISE 是基于 WISE 巡天图像数据重新处理发布的星表，包括了近 7 亿个源在中红外波段的测光数据。相比 WISE 数据星表，ALLWISE 具有更高的测光灵敏度、准确性和天测精度。AllWISE 数据集包括了 WISE 观测四个阶段的数据，第一阶段是红外巡天观测，从 2010 年 1 月持续到 2010 年 9 月，共 4 个波段（$W1$、$W2$、$W3$、$W4$); 第二阶段是在望远镜冷却剂耗尽以后进行的为期四个月的近地小行星观测，称之为 NEOWISE，只有 $W1$、$W2$ 两个波段。考虑到 $W3$ 和 $W4$ 波段的星等误差较大，在本研究中只使用了 $W1$、$W2$ 波段的测光数据。各个数据集的下载链接列在表 3-1 中。

然后，我们开始已知样本的构建，主要包括如下几个步骤：

(1) 从 BASS DR3 数据中选择符合条件的数据，包括星等范围约束：$0 < gPSFMag \leq 24.2$, $0 < rPSFMag \leq 23.6$, $0 < zPSFMag \leq 23$; 数据质量约束：Flag_ISO_g= 0, Flag_Model_g= 0, Flag_ISO_r= 0, Flag_Model_r= 0, Flag_ISO_z= 0, and Flag_Model_z= 0, 这些字段分别表示孔径及模型星等在 g、r、z 波段的质量，值大于 0 时，表示有不同的质量问题，因此我们要求值等于 0。最终选择的数据总量为 110,896,598 个天体；

(2) 参考 Schlegel 等人的方法 (Schlegel et al., 1998) 对上述数据进行红化校正，g、r、z 波段的校正因子分别为：3.303、2.285、1.263；

(3) 将 SDSS DR16 中 zWarning=0，类别为恒星和星系的数据，SDSS DR16Q 中 zWarning=0 的类星体数据，LAMOST DR5 中的星系、类星体及信噪比（S/N）

大于 10 的恒星分别与前一步骤中的数据按位置交叉，交叉半径为 2 角秒，同一个源的交叉结果保留距离最近的一个，最终将交叉后的结果合并，称为 BASS-SDSS-LAMOST 样本。如果同一个源在 SDSS 和 LAMOST 中都有对应的光谱，则只保留 SDSS 中的数据, 我们将此样本记为 SAMPLE I；

(4) 将 BASS-SDSS-LAMOST 样本与 ALLWISE 数据集交叉，交叉半径为 4 角秒，获取对应的 $W1$、$W2$ 波段的数据，最终形成训练与测试样本集 BASS-SDSS-LAMOST-ALLWISE，由于 ALLWISE 中默认采用的是 VEGA 星等，需要将星等转换为 AB 星等；

(5) 对 BASS-SDSS-LAMOST-ALLWISE 样本中的 $W1$、$W2$ 星等进行红化校正，校正因子分别为：0.189、0.146。我们将此样本记为 SAMPLE II。最终样本中的主要字段信息详见表 3-2。

表 3-2 已知样本中主要字段说明。

Table 3-2 The columns, definition, catalogues, and wavebands.

字段名称	字段描述	来源星表	波段
id	源 ID	BASS	
ra	赤经	BASS	
dec	赤纬	BASS	
$gKronMag$	g 波段 Kron 星等	BASS	光学波段
$rKronMag$	r 波段 Kron 星等	BASS	光学波段
$zKronMag$	z 波段 Kron 星等	BASS	光学波段
$gPSFMag$	g 波段 PSF 星等	BASS	光学波段
$rPSFMag$	r 波段 PSF 星等	BASS	光学波段
$zPSFMag$	z 波段 PSF 星等	BASS	光学波段
$W1mag$	$W1$ 波段星等	ALLWISE	红外波段
$W2mag$	$W2$ 波段星等	ALLWISE	红外波段
CLASS	类型标签	SDSS, LAMOST	
$Redshift$	光谱红移	SDSS, LAMOST	
g	消光校正后的 g 波段星等	BASS	光学波段
r	消光校正后的 r 波段星等	BASS	光学波段
z	消光校正后的 z 波段星等	BASS	光学波段
$W1$	消光校正后的 $W1$ 波段星等	ALLWISE	红外波段
$W2$	消光校正后的 $W2$ 波段星等	ALLWISE	红外波段

3.2　基于星等-颜色的分类

根据点源与展源的观测特征，Kron 星等适合星系，而 PSF 星等更适合恒星和类星体。在同一波段的 Kron 星等与 PSF 星等的差，展源通常要大于点源。金鑫等 (Jin et al., 2019) 在对 PanSTARRS 数据进行分类时，通过计算 $iPSFMag - iKronMag$ 和 $zPSFMag - zKronMag$ 分布，能明显区分点源与展源，准确率达到了 96% 以上。由此可见，同一波段 Kron 星等与 PSF 星等的差可以大致区分点源和展源，对于本实验样本是否具有相同的特性，我们先做类似的统计。为了便

于描述，我们定义 $\Delta g = gPSFMag - gKronMag$, $\Delta r = rPSFMag - rKronMag$, $\Delta z = zPSFMag - zKronMag$。训练样本 Δg、Δr、Δz 的分布如图 3-1 所示。

由图 3-1 可见，Kron 星等与 PSF 星等之差能够很大程度上进行点源与展源的区分，当选择 $\Delta r > 0.20$ 或者 $\Delta z > 0.20$ 时，91.0% 的星系能够被排除。然而，如图 3-1 左侧部分可见，随着天体变暗，星系与恒星、类星体越来越难以区分。Yang 等 (Yang et al., 2017) 的工作也指出，测光数据的形态学特征，在暗端区分能力较差。而且，对于同为点源的恒星与类星体，则基本上都完全重合在一起。因此，需要寻找更为有效的分类方法和分类特征。为了进一步分析颜色特征对于分类的影响，我们将不同颜色分布进行一个 2 维空间的可视化（图 3-2、图 3-3），从而发现双色图对于星系、恒星与类星体之间的可分辩性。

从图 3-2 和图 3-3 可知，对于任何的单个特征或两个特征都很难将恒星与类星体进行有效区分。但相比较图 3-3 与图 3-2，红外特征对于类星体与恒星的分类具有一定的区分度，这也进一步表明，红外特征对于类星体的发现与认证具有很大的贡献，这与类星体的特征相一致。根据目前对类星体的了解，类星体具有宽波段能谱分布，在射电、红外、可见光、紫外、X 射线和 γ 射线波段均具有明显的特征，而在可见光图像上与恒星非常相似。因此，我们基于多波段星等及颜色特征，在更高维的空间上采用机器学习算法来区分恒星和类星体会更有效。

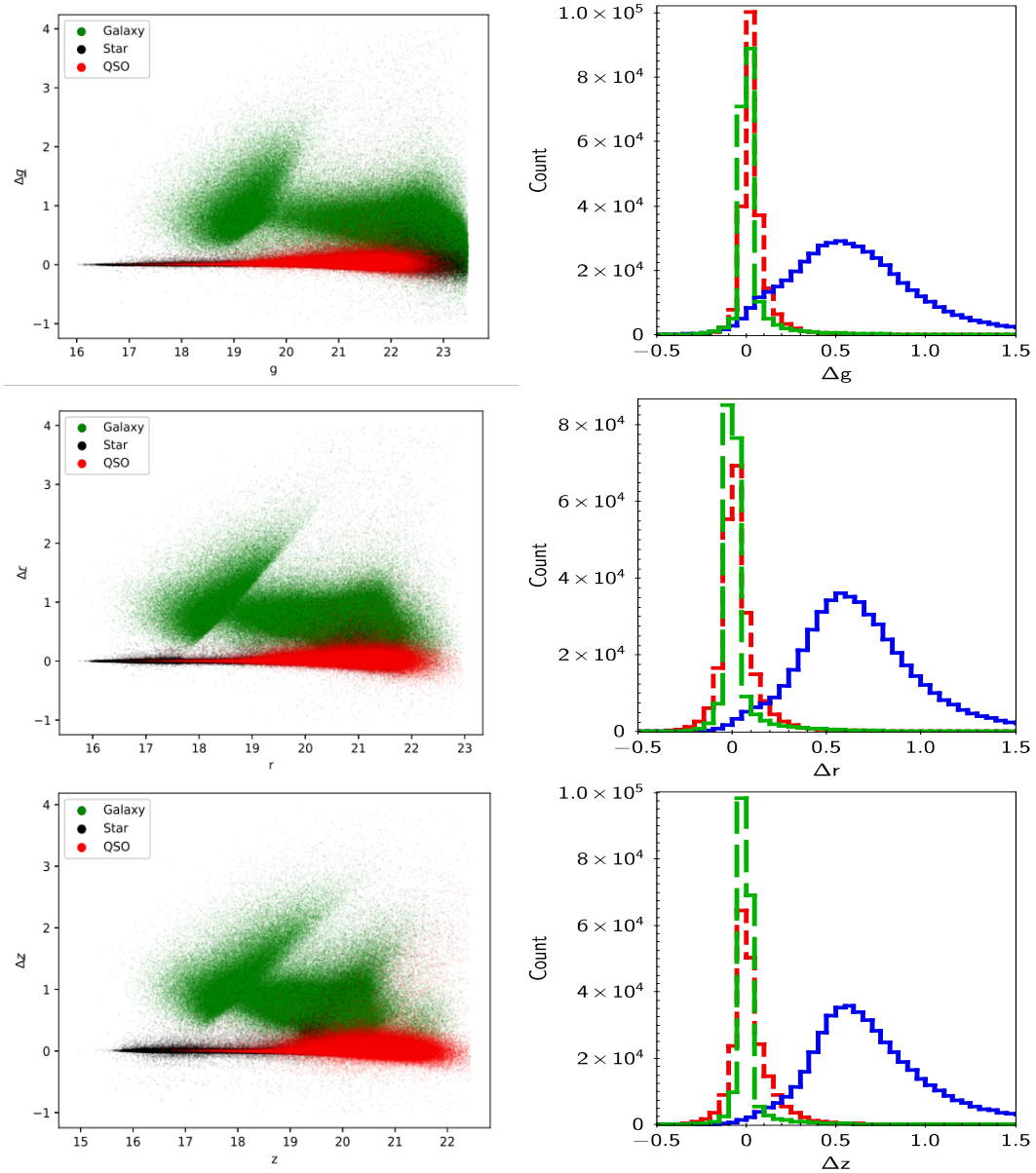


图 3-1 左侧部分为星系、恒星、类星体的 Δg, Δr, Δz 与对应星等的分布，绿色表示星系、红色表示类星体、黑色表示恒星。右侧部分为 Δg, Δr, Δz 的区间分布直方图，绿色为星系、红色为类星体、蓝色为恒星。

Figure 3-1 Panel Left: Panel Right: The distribution of Δg, Δr, Δz of known stars, quasars, and galaxies. The green long dash line represent stars, the blue line represents galaxies, and the red dash line represents quasars.

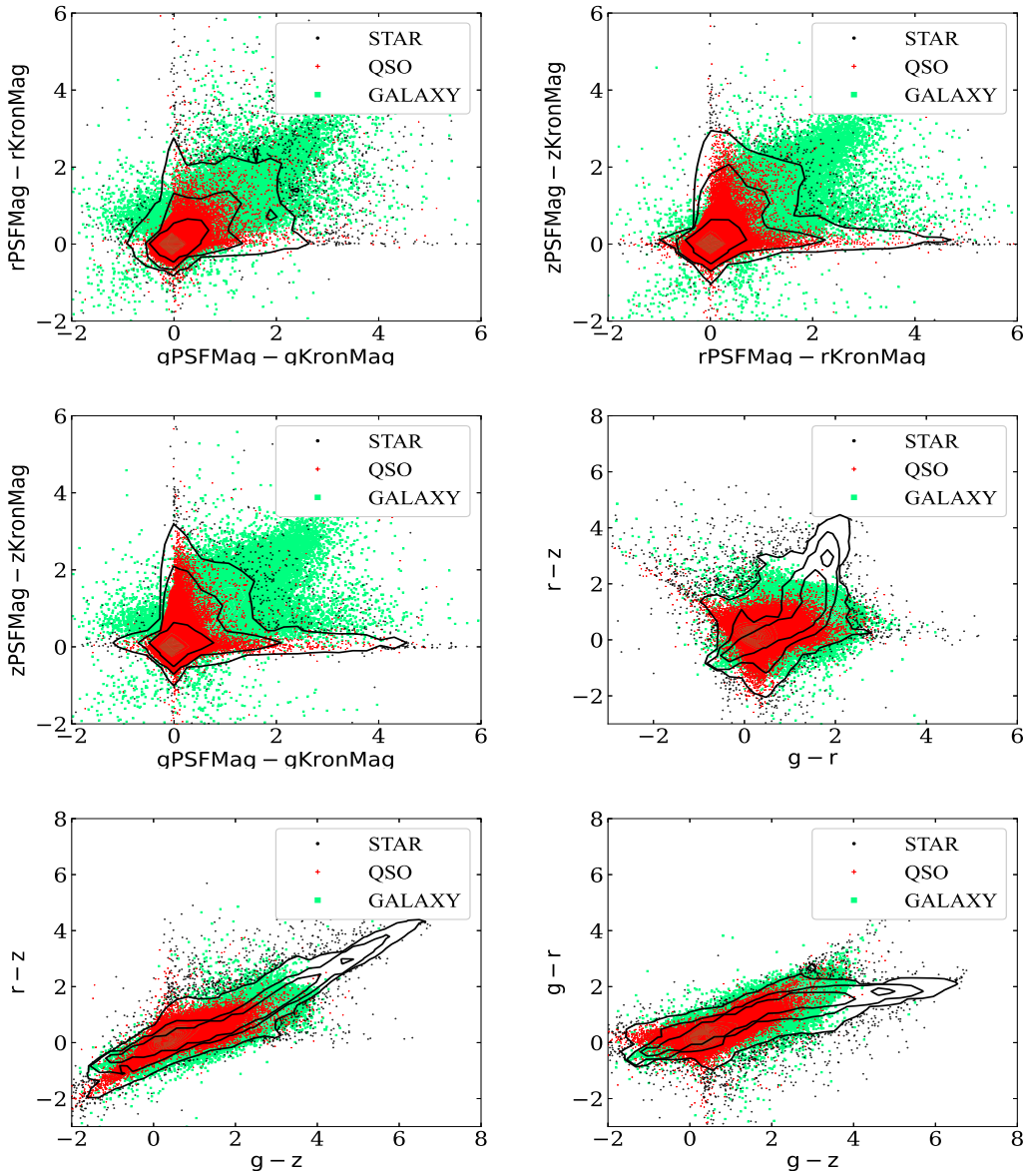


图 3-2 星系、恒星及类星体在 2 维可见光颜色空间上的分布，绿色表示星系、红色表示类星体、黑色表示恒星，黑色的等密度线表示恒星的密集分布区域。

Figure 3-2 The distribution of stars, galaxies and quasars in 2-d optical spaces, green filled squares represent galaxies, red pluses represent quasars and black filled circles represent stars, the black outline is the contour of star distribution.

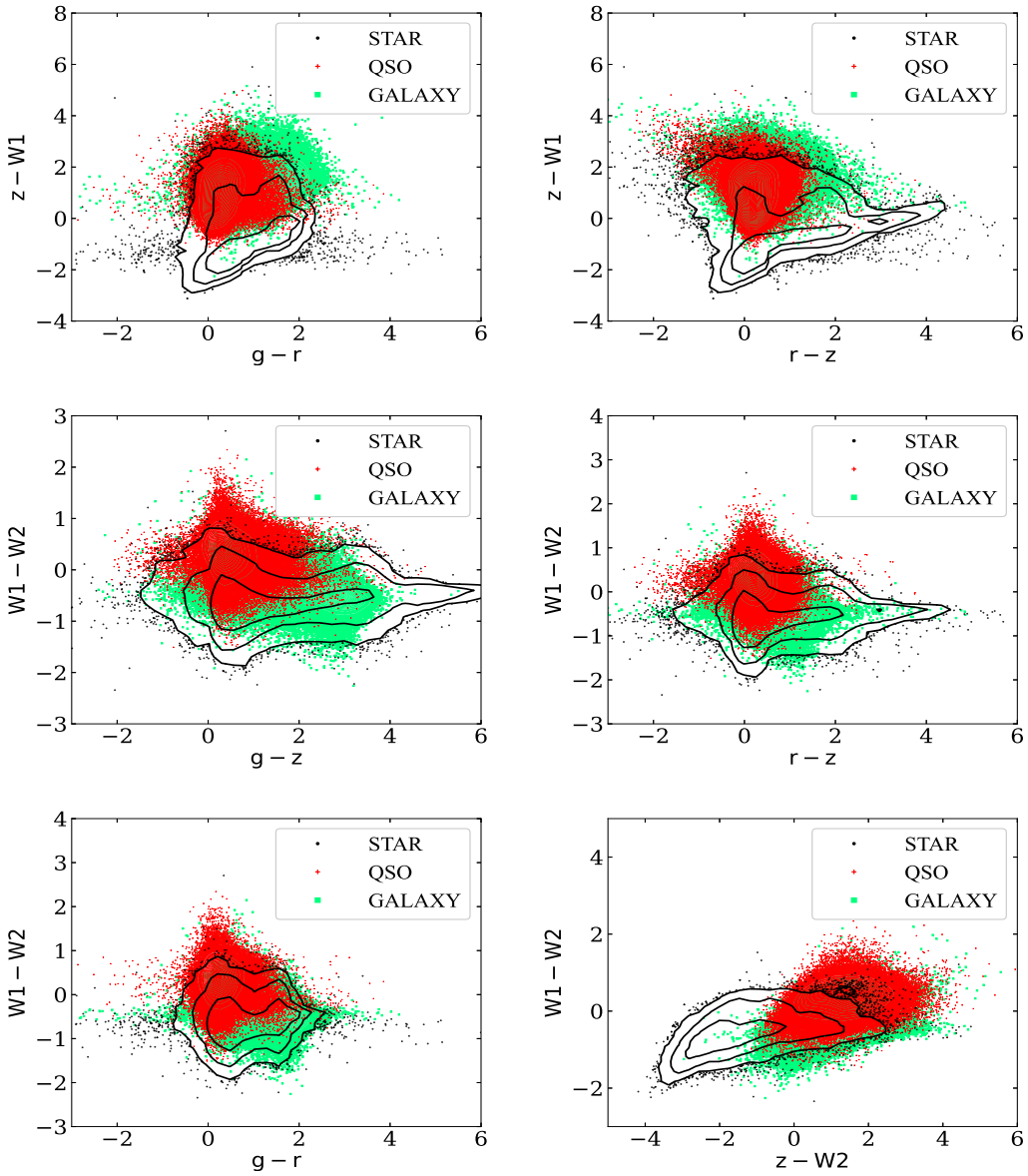


图 3-3 星系、恒星及类星体在 2 维红外相关波段颜色空间上的分布，绿色表示星系，红色表示类星体，黑色表示恒星，黑色的等密度线表示恒星的密集分布区域。

Figure 3-3 The distribution of stars, galaxies and quasars in 2-d infrared spaces, green filled squares represent galaxies, red pluses represent quasars and black filled circles represent stars, the black outline is the contour of star distribution.

3.3 机器学习算法及评价指标

在算法选择上，我们选择 XGBoost 作为主要算法，同时与随机森林算法进行了对比。为了描述分类的性能，我们先定义评价性能的主要指标。对于分类算法，通常采用的评价指标有准确率（Accuracy)、精度（Precision)、召回率（Recall)和 F1_Score (F1) 等。准确率是指对于分类样本中所有预测正确的数量与该数据样本总数量的比例。以正类样本为例，精度（又称查准率）是指预测正确的正类数量占预测为正类样本的总数的比例；召回率（又称查全率、灵敏度）是指预测正确的正类数量占实际为正类样本总数的比例；F1 表示精度与召回率之间的一种加权平均。对于二元分类器，精度和召回率出现矛盾时，可以利用 F1 来综合评价，其值越大越好。为了便于理解计算，通常采用混淆矩阵（Confusion matrix)的方式来描述。对于二类的情况，准确率、精度、召回率的计算方法分别如公式 3-1、3-2、3-3 和 3-4 所示，其中 TP 和 TN 分别表示被分类器正确分类的正类和负类的数量，FP 和 FN 表示被错误分类的正类和负类的数量：

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3-1}$$

$$\text{Precison (Prec.)} = \frac{TP}{TP + FP} \tag{3-2}$$

$$\text{Recall (Rec.)} = \frac{TP}{TP + FN} \tag{3-3}$$

$$\text{F1\_score (F1)} = \frac{2 \times (\text{Prec.} \times \text{Rec.})}{\text{Prec.} + \text{Rec.}} \tag{3-4}$$

为了便于更好地理解三类分类情况下的指标计算方法，表 3-3 给出了星系、恒星、类星体三类情况下的混淆矩阵表示，对应每一类分类的精度、召回率以及准确率的计算方法。表 3-3 中，TG 表示预测与真值都是星系的数量；FGQ 表示被误分为类星体的星系的数量；FGS 表示被误分为恒星的星系的数量；TQ 表示预测与真值都是类星体的数量；FQG 表示被误分为星系的类星体的数量；FQS 表示被误分为恒星的类星体的数量；TS 表示预测与真值都是恒星的数量；FSG 表示被误分为星系的恒星的数量；FSQ 表示被误分为类星体的恒星的数量。

表 3-3 三类分类时的混淆矩阵。

Table 3-3 Confusion matrix for three-class classification.

已知↓预测→	星系	类星体	恒星	精度	召回率
星系	TG	FGQ	FGS	$\frac{TG}{TG+FQG+FSG}$	$\frac{TG}{TG+FGQ+FGS}$
类星体	FQG	TQ	FQS	$\frac{TQ}{TQ+FGQ+FSQ}$	$\frac{TQ}{TQ+FQG+FQS}$
恒星	FSG	FSQ	TS	$\frac{TS}{TS+FGS+FQS}$	$\frac{TS}{TS+FSG+FSQ}$
准确率	$\frac{TG+TQ+TS}{TG+FGQ+FGS+TQ+FQG+FQS+TS+FSG+FSQ}$				

3.4　分类器构建

3.4.1　最优特征选择

传统机器学习算法的输入是一组向量特征，组成向量的特征个数及每个特征的重要性对于算法的性能具有较大的影响。通常认为，特征越多，性能越好，但有时也会出现特征多反而性能下降的情况。为了找到最优的特征组合，我们需要对特征的重要性进行评估，越重要的特征，放在输入向量的前面。在第二节简单对比了不同特征空间对于分类的影响，而算法 XGBoost 本身也具备特征评估能力，可以给出每个特征重要性的详细得分。图 3-4 给出了四种输入特征分类情况下特征重要性的排序。

从图 3-4 可见，对 SAMPLE I 样本，在只采用可见光特征时，分类点源与展源的特征重要性排序从大到小为 Δz, Δr, Δg, $g-r$, g, $g-z$, r, z, $r-z$; 而区分类星体与恒星时，排序为 r, $g-z$, z, $g-r$, Δz, $r-z$, Δr, Δg, g。对 SAMPLE II 样本而言，考虑可见光与中红外特征，分类点源与展源时，特征重要性排序从大到小为 Δg, Δz, $g-W1$, $W1-W2$, $z-W1$, Δr, $g-z$, $z-W2$, $g-r$, $r-z$, $W1$, r, g, z, $r-W2$, $W2$, $r-W1$, $g-W2$；区分类星体与恒星时的特征排序为 $z-W2$, $W1-W2$, $g-z$, $g-r$, $z-W1$, $r-z$, Δz, r, $r-W2$, z, Δg, $g-W1$, Δr, $W1$, $g-W2$, g, $r-W1$, $W2$。图 3-3 和 3-4 也表明，特征重要性与分类任务、样本数据紧密相关。

3.4.2　XGBoost 二元分类器构建

一个二元分类器只能进行两种类别的区分。因此，我们对于星系、恒星、类星体的区分可以分两步来完成，即先进行点源（恒星和类星体）与展源（星系）的分类，然后再对点源进行恒星与类星体的分类。根据上一节的特征重要性排序，我们分别进行了二元分类器的最优模型参数的选择。我们采用类似网格搜索（grid search）的参数最优化方法和 10 折交叉验证的评价策略，对 XGBoost 的主要超参数进行最优选择，在 $max\_depth = [5, 7, 9, 11, 13]$, $n\_estimators = [100, 200, 300]$

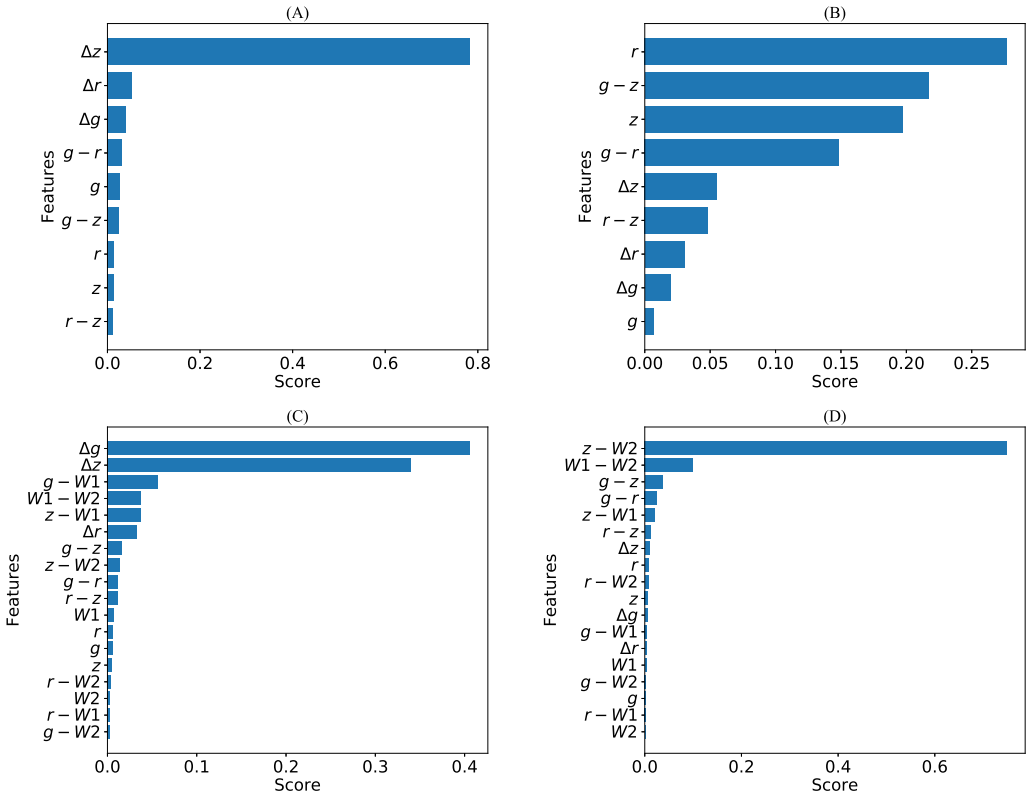


图 3-4 针对样本 I 和样本 II，XGBoost 算法给出的特征重要性排序。图 (A) 为只采用光学波段特征分类点源与展源; 图 (B) 为只采用光学波段特征分类恒星与类星体；图 (C) 为采用光学与红外特征分类点源与展源；图 (D) 为采用光学与红外特征分类恒星与类星体。

Figure 3-4 The feature importance rank with XGBoost. Panel(A) shows the feature importance only with the optical information when classifying extended sources and point sources; panel(B) shows the feature importance only with the optical information when classifying stars and quasars; panel(C) shows the feature importance with combined optical and infrared information when classifying extended sources and point sources; panel(D) shows the feature importance with combined optical and infrared information when classifying stars and quasars.

和 $learning\_rate = [0.1, 0.3, 0.5]$ 等超参数空间上进行训练，得到的最好模型超参数为 $max\_depth = 11$, $n\_estimators = 100$ 和 $learning\_rate = 0.5$。然后，我们对输入特征进行多种组合训练，表 3-4 列出了不同输入特征时，分类点源与展源分类器获得的准确率、精度、召回率及训练时间。

表 3-4 表明，在只有光学特征时，最好的准确率为 97.28%; 采用光学与红外组合特征时，最好的准确率达到 98.67%, 对于点源的精度与召回率则都达到了 98.30%，而对于展源的精度与召回率则都超过了 98.80%，远高于基于颜色的截断获得的分类性能。因此，可以认为在光学与红外特征相结合时，XGBoost 算法对于点源与展源的分类是高效的、可靠的。

然后，我们同样在 SAMPLE I 和 SAMPLE II 上进行恒星与类星体的分类训练。表 3-5 列出了不同输入特征时，分类恒星与类星体时的准确率、精度、召回

表 3-4 分类点源与展源的二元分类器的准确率、精度及召回率。

Table 3-4 The performance of binary classifier for point and extended sources.

		点源			展源		
输入特征	训练样本	准确率 (%)	精度.(%)	召回率 (%)	精度 (%)	召回率 (%)	Time(s)
(Δg, Δr, Δz)	SAMPLE I	95.72	94.98	95.75	96.36	95.69	21
(Δg, Δr, Δz, $g-r$, $r-z$)	SAMPLE I	96.73	96.07	96.84	97.29	96.63	34
(Δg, Δr, Δz, $g-r$, $r-z$, $g-z$)	SAMPLE I	96.78	96.08	96.96	97.39	96.63	47
(Δg, Δr, Δz, $g-r$, $r-z$, r)	SAMPLE I	97.22	96.57	97.40	97.77	97.06	54
(8-features)	SAMPLE I	**97.28**	**96.63**	**97.47**	**97.83**	**97.11**	73
(11-features)	SAMPLE II	98.52	98.26	98.11	98.70	98.80	90
(13-features)	SAMPLE II	98.64	98.35	98.32	98.84	98.86	105
(15-features)	SAMPLE II	**98.67**	**98.39**	**98.35**	**98.86**	**98.89**	111
(18-features)	SAMPLE II	98.65	98.37	98.32	98.84	98.87	152

[a] 8-features 代表 Δg, Δr, Δz, $g-r$, $g-z$, r, $r-z$, g，共八个特征。

[b] 11-features 代表 Δg, Δr, Δz, $g-r$, $r-z$, $g-z$, g, r, z, $W1$, $W2$，共十一个特征。

[c] 13-features 代表 Δg, Δz, $g-W1$, $W1-W2$, $z-W1$, Δr, $g-z$, $z-W2$, $g-r$,$r-z$,$W1$, r, g，共十三个特征。

[d] 15-features 代表 Δg, Δz, $g-W1$, $W1-W2$, $z-W1$, Δr, $g-z$, $z-W2$, $g-r$,$r-z$,$W1$, r, g, z, $r-W2$，共十五个特征。

率及训练时间。

正如表 3-5 所述，只采用光学特征时，恒星与类星体的分类最好情况下的准确率为 93.22%，恒星的分类精度与召回率分别为 93.33% 和 93.71%, 类星体的分类精度与召回率则分别为 93.11% 和 92.69%；而当同时考虑光学与红外特征时，最好情况下的准确率达到 99.15%，恒星的分类精度与召回率分别为 99.00% 和 99.46%，类星体的分类精度与召回率分别为 99.33% 和 98.77%，分类性能明显提升。分类结果进一步表明，红外特征对于恒星与类星体的区分具有重要的影响，这与特征重要性的评估结果及对类星体的现有认识是一致的。

当使用两步的二元分类器时，第一步的分类性能会直接影响第二步的分类效果。因此，总体性能可以简单认为是两步性能的乘积，即用二步分类器完成星系、恒星、类星体的分类时，最好的准确率为 97.83%。

3.4.3 XGBoost 多元分类器构建

相比两步的二元分类器，多元分类器使用起来则更为简单，XGBoost 可以通过调参直接支持。我们只需要将 XGBoost 模型训练的参数 *objective* 设置为 $multi:softmax$ 及 num_class=3。然后我们同样采用网格搜索和 10 折交叉验证的方法，得到的最优模型超参数为 $max\_depth = 7$, $n\_estimators = 200$ 和 $learning\_rate = 0.5$。表 3-6 列出了多元分类器下不同输入特征时所训练的分类器性能，包括两组只有光学特征及两组光学与红外相结合的特征。当只使用光学特征时，最好的分类准确率为 94.49%，而当使用光学与红外相结合的特征时，最好的分类准确率达到了 98.43%，分类类星体的精度与召回率则分别为 97.07% 和 97.95%。

3.4.4 与随机森林性能对比

除了 XGBoost，我们也采用随机森林方法进行了恒星与类星体的分类实验，训练的过程与 XGBoost 基本一致。我们同样进行了超参数的最优化及输入特征的最优化，表 3-7 列出了在 SAMPLE I 和 SAMPLE II 上的分类性能。如果只考虑光学特征，最好的性能为准确率 93.27%，而在考虑光学与红外特征的情况下，准确率达到了 99.14%。虽然准确性与 XGBoost 算法很接近，但在相同的分类目标与样本上，XGBoost 的计算速度比随机森林快很多倍，这进一步表明 XGBoost 算法是适合于大样本训练的学习方法。

3.4.5 讨论

我们的研究目的是要利用 XGBoost 方法从测光数据中对星系、恒星、类星体进行区分，采用的测光信息为三个光学波段（g、r、z）及两个中红外波段（$W1$、$W2$）。正如图 3-1 所示，从数据中直接区分星系还是比较容易的。由于星系的展源特征，Δg、Δr 和 Δz 的值都不等于 0，而恒星与类星体的 Δg、Δr 和 Δz 的绝对值却非常接近 0。因此，Δg、Δr 和 Δz 也成为了 XGBoost 分类时的重要输入特征。同时，XGBoost 也结合了其它特征的一些细微差异进行综合考虑，分类性

表 3-5 分类恒星和类星体的 XGBoost 二元分类器的准确率、精度及召回率。

Table 3-5 The performance of binary classifier for stars and quasars by XGBoost.

			恒星		类星体		
输入特征	训练样本	准确率 (%)	精度 (%)	召回率 (%)	精度 (%)	召回率 (%)	时间 (s)
(Δg, Δr, Δz, $g-r$)	SAMPLE I	84.68	83.13	88.62	86.62	80.38	12
(r, $g-z$,z,$g-r$)	SAMPLE I	91.31	91.40	91.99	91.20	90.56	17
(r, $g-r$,$r-z$)	SAMPLE I	91.75	91.71	92.55	91.79	90.87	12
(Δg, Δr, Δz, $g-r$, $r-z$, r)	SAMPLE I	93.20	93.25	**93.75**	**93.14**	92.60	25
(9-features)	SAMPLE I	**93.22**	**93.33**	93.71	93.11	**92.69**	37
($g-r$, $r-z$, r, g, z, $w1$, $w2$)	SAMPLE II	98.88	98.61	99.38	99.23	98.27	24
(10-features)	SAMPLE II	99.11	98.95	99.45	99.32	98.70	29
(12-features)	SAMPLE II	99.13	98.92	99.51	99.40	98.67	17
(15-features)	SAMPLE II	**99.15**	**99.00**	99.46	99.33	**98.77**	25
(17-features)	SAMPLE II	99.12	98.94	**99.47**	**99.35**	98.69	36

[a] 9-features 代表 Δg, Δr, Δz, $g-r$, $r-z$, $g-z$, r, g, z，共 9 个特征。

[b] 10-features 代表 $z-W2$,$z-W1$, $g-r$, $W1-W2$, Δz,Δg, Δr, $r-z$, $r-W2$, r，共 10 个特征。

[c] 12-features 代表 $z-W2$,$z-W1$, $g-z$, $W1-W2$, $g-r$, Δz, $r-z$, $r-W2$, r, z, Δg, $g-W1$，共 12 个特征。

[d] 15-features 代表 $z-W2$, $W1-W2$, $g-z$, $g-r$, $z-W1$, $r-z$, Δz, r, $r-W2$, z, Δg, $g-W1$, Δr, $W1$, $g-W2$，共 15 个特征。

[e] 17-features 代表 $z-W2$,$z-W1$, $g-z$, $W1-W2$, $g-r$, Δz, $r-z$, $r-W2$, r, z, Δg, $g-W1$, Δr, g, $g-W2$, $W1$, $W2$，共 17 个特征。

表 3-6 不同输入特征下的 XGBoost 多元分类器的性能。

Table 3-6 The performance of multiclass classifier for galaxies, stars, and quasars by XGBoost.

			星系		恒星		类星体		
输入特征	训练样本	准确率 (%)	精度 (%)	召回率 (%)	精度 (%)	召回率 (%)	精度 (%)	召回率 (%)	训练时间 (s)
Input pattern I	SAMPLE I	94.47	97.74	97.32	92.96	90.59	**88.68**	91.63	326
Input pattern II	SAMPLE I	**94.49**	**97.76**	**97.34**	**93.00**	**90.60**	88.67	**91.65**	260
Input pattern III	SAMPLE II	97.90	98.36	98.76	98.32	95.50	96.37	97.62	160
Input pattern IV	SAMPLE II	**98.43**	**98.85**	**98.97**	**98.74**	**97.28**	**97.07**	**97.95**	330

[a] Input pattern I 代表 Δg, Δr, Δz, $g-r$, $r-z$, r，共 6 个特征。

[b] Input pattern II 代表 Δg, Δr, Δz, $g-r$, $g-z$, r, $r-z$, z，共 8 个特征。

[c] Input pattern III 代表 $z-W2$, Δz, $W1-W2$, Δr, $g-r$, $z-W1$, $W1$, $W2$，共 8 个特征

[d] Input pattern IV 代表 $z-W2$, Δz, $W1-W2$, Δr, $g-r$, $z-W1$, Δg, $g-z$, $r-W2$, $r-z$, r，共 11 个特征。

表 3-7 随机森林进行恒星与类星体分类时的性能。

Table 3-7 The performance of binary classifier for stars and quasars by random forest.

			STAR		QSO		
输入特征	样本	准确率 (%)	精度 (%)	召回率 (%)	精度 (%)	召回率 (%)	训练时间 (s)
(Δg, Δr, Δz, $g-r$)	SAMPLE I	85.51	83.52	89.99	88.07	80.63	125
(Δg, Δr, Δz, $g-r$, $r-z$, r)	SAMPLE I	93.17	92.31	**94.81**	**94.16**	91.39	120
(**9-features**)	SAMPLE I	**93.27**	**92.83**	94.39	93.77	**92.05**	180
($g-r$, $r-z$, r, g, z, $w1$, $w2$)	SAMPLE II	98.77	98.24	99.55	99.44	97.81	80
(15-features)	SAMPLE II	**99.14**	**98.82**	**99.62**	**99.53**	**98.54**	120

[a] 9-features 代表 Δg, Δr, Δz, $g-r$, $r-z$, $g-z$, r, g, z，共 9 个特征。

[b] 12-features 代表 $z-W2$,$z-W1$, $g-z$, $W1-W2$, $g-r$, Δz, $r-z$, $r-W2$, r, z, Δg, $g-W1$，共 12 个特征。

[c] 15-features 代表 $z-W2$,$W1-W2$,$g-z$,$g-r$,$z-W1$,$r-z$,Δz,r,$r-W2$,z,Δg,$g-W1$,Δr,$W1$,$g-W2$，共 15 个特征。

能要远高于单独的星等截断方法获得的性能，这与表 3-4 到表 3-6 的实验结果是相一致的。虽然采用同样的算法来训练二元分类器与多元分类器，但二元与多元分类器在树模型的构造上具有显著的差别。如果同一个源在两个分类器中所分的类型相一致，那么可以认为这个分类是更可靠的。因此，采用两种分类策略相互间也是一个验证。通过对比表 3-4、3-5 与表 3-6，我们发现当采用两步的二元分类器时，恒星与类星体的精度与召回率都要高于采用多元分类器，不论样本是 SAMPLE I 还是 SAMPLE II。由此表明，对于重在针对甄选类星体的任务，采用二元分类器的结果会更可靠些。

在 Stern 等 (Stern et al., 2005) 和 Hickboxt 等 (Hickox et al., 2007) 的研究中表明，红外信息对于区分类星体是非常有效的。在 Bovy(Bovy et al., 2012)、DiPompeo(DiPompeo et al., 2015) 等的搜寻类星体候选体的工作中，同样也表明了红外信息的重要性。我们的实验进一步证明，红外信息对于从测光数据中区分类星体具有非常重要的作用。

在分类算法选择上，我们在采用 XGBoost 作为主分类器时，与使用随机森林进行恒星与类星体分类的性能作了比较。表 3-7 列出的采用随机森林时的分类性能。对于 SAMPLE I 样本，最好的性能为准确率 93.27%，对于 SAMPLE II 样本，最好的性能为准确率 99.14%，类星体的精度为 99.53%，召回率为 98.54%。将表 3-7 与表 3-5 比较发现，虽然两种算法在性能上相当，但 XGBoost 的训练时间明显更短。由此表明，对于本次实验样本，XGBoost 要优于随机森林，而且也表明，XGBoost 在大数据时代也更具有应用潜力。

3.5 BASS DR3 数据的分类

根据前面的实验及其结果对比，我们采用 XGBoost 算法、最优的输入特征、最优的模型参数构建了六个分类模型，表 3-8 列出了这些模型的详细信息，六个分类器模型分别对六种不同的分类目标及样本特征。如果采用二元分类模型，需要二个分类器来完成星系、恒星及类星体的分类，而如果采用多元分类器，则只需要一个模型就可以完成。同时，考虑到 BASS DR3 中包括大量未找到红外波段信息的天体，我们也提供仅包括光学输入特征的分类器。这些分类器也适用于对类似观测数据的不同样本进行预测。基于六个分类器进行 BASS DR3 天体分类的流程如图 3-5 所示，图中红框表示数据分析过程，而黑色平行四边形表示数据的合并或结果。

由于 BASS DR3 数据本身只有 g、r、z 三个光学信息，按照图 3-5 所示流程，我们使用分类器 I 将 BASS DR3 (110,896,598) 的源分成展源与点源，然后再使用分类器 II 将点源分类成恒星与类星体，同时也使用分类器 III 将 BASS DR3 的源直接分成恒星、类星体与星系。然后我们将 BASS DR3 的源与 ALLWISE 交叉，得到具有红外特征的数据集，表示为 BASS-DR3-ALLWISE，共有 43,859,467 个源。我们使用分类器 IV 将 BASS-DR3-ALLWISE 进行展源与点源的分类，再使用分类器 V 进行恒星与类星体的分类，同时也采用分类器 VI 直接将 BASS-DR3-

表 3-8 针对不同目标构建的六个 XGBoost 分类器模型。

Table 3-8 The six XGBoost classifiers for different tasks.

分类器	分类目标	输入特征	分类类型
Classifier 1^{st}	binary:logistic	(Δg, Δr, Δz, $g-r$, $g-z$, r, $r-z$, g)	(点源与展源)
Classifier 2^{nd}	binary:logistic	(Δg, Δr, Δz, $g-r$, $r-z$, $g-z$, r, g, z)	(恒星与类星体)
Classifier 3^{rd}	multi:softmax	(Δg, Δr, Δz, $g-r$, $g-z$, r, $r-z$, z)	(星系、恒星及类星体)
Classifier 4^{th}	binary:logistic	(Pattern I)	(点源与展源)
Classifier 5^{th}	binary:logistic	(Pattern II)	(恒星与类星体)
Classifier 6^{th}	multi:softmax	(Pattern III)	(星系、恒星及类星体)

[a] 展源指星系，而点源指恒星和类星体。

[b] Pattern I 代表 Δg, Δz, $g-W1$, $W1-W2$, $z-W1$, Δr, $g-z$, $z-W2$, $g-r$,$r-z$,$W1$, r, g, z, $r-W2$.

[c] Pattern II 代表 $z-W2$,$W1-W2$,$g-z$,$g-r$,$z-W1$,$r-z$,Δz,r,$r-W2$,z,Δg,$g-W1$,Δr,$W1$,$g-W2$.

[d] Pattern III 代表 $z-W2$, Δz, $W1-W2$, Δr, $g-r$, $z-W1$, Δg, $g-z$, $r-W2$, $r-z$, r.

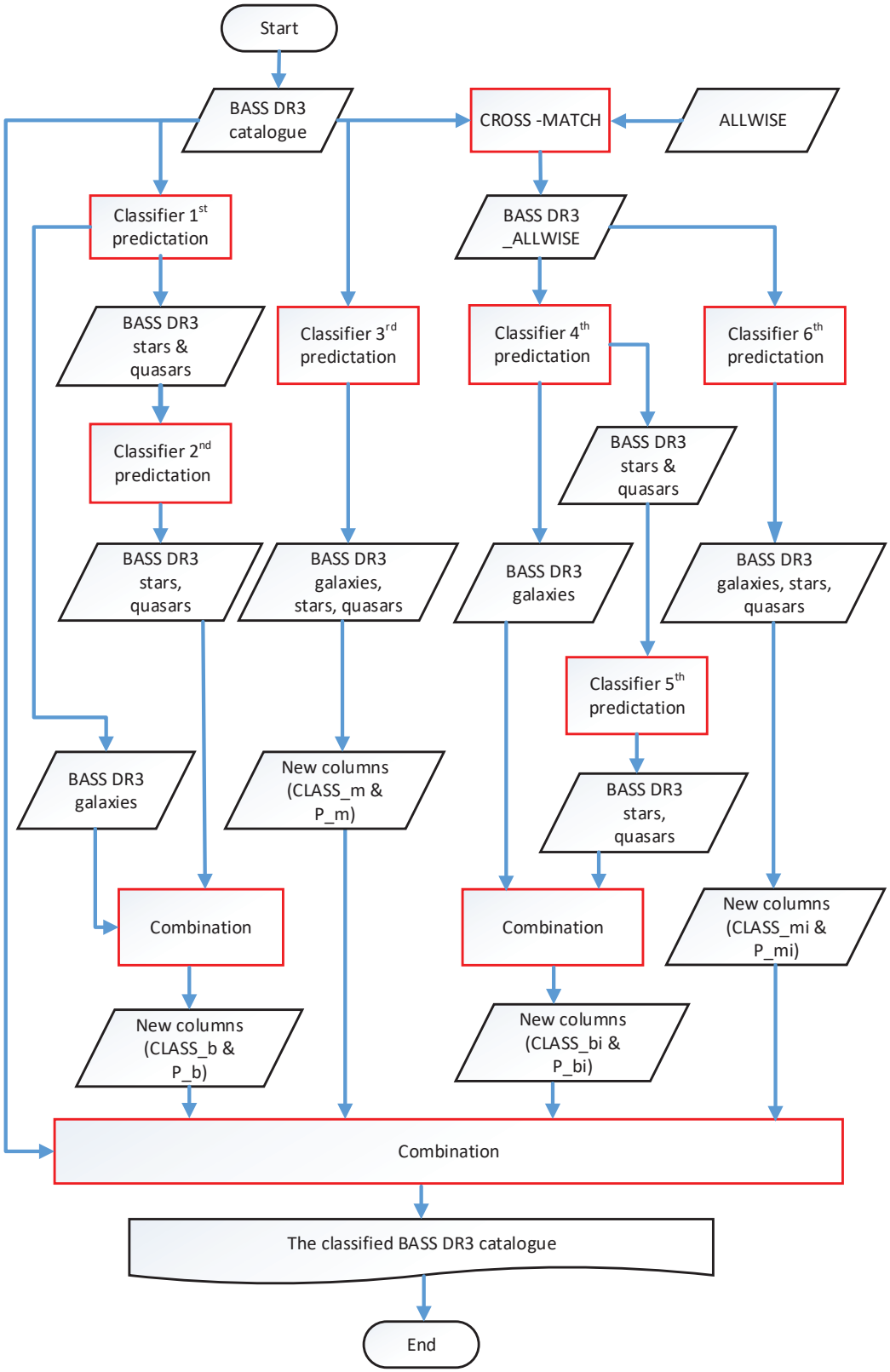


图 3-5 对 BASS DR3 天体进行分类预测的流程图。

Figure 3-5 The classification workflow.

ALLWISE 分类成星系、恒星和类星体。每次分类的结果包括分类标签及置信度。最后我们将所有分类的结果按照天体源的 ID 进行合并，形成新的星表，星表的数据结构如表 3-9 所示，全部预测结果下载链接为https://doi.org/10.12149/101065。分类结果可用于对天体后续进一步的研究及作为光谱巡天的输入星表进行类星体证认。

我们对 BASS DR3 天体分类的结果进行了进一步的统计分析。表 3-10 列出了使用不同分类器时，分类结果中星系、恒星和类星体的数量。如果只采用光学特征，二元及多元分类器预测结果相同且置信度大于 75% 的星系、恒星和类星体的数量分别为 19,829,533 ($P_S > 0.75$)、49,483,839 ($P_G > 0.75$) 和 7,095,580 ($P_Q > 0.75$)。其中置信度达到 90% 的类星体数量为 2,775,970，置信度达到 95% 的类星体数量达到了 1,166,517。而如果同时考虑光学与红外特征，二元及多元分类器预测结果相同时，置信度大于 75%、90%、95% 的恒星数量分别为 12,785,232 ($P_S > 0.75$)、12,561,500 ($P_S > 0.90$) 和 12,375,838 ($P_S > 0.95$); 置信度大于 75%、90%、95% 的星系数量分别为 25,068,898 ($P_G > 0.75$)、21,890,547($P_G > 0.90$) 和 18,606,073 ($P_G > 0.95$); 置信度大于 75%、90%、95% 的类星体数量分别为 1,500,099 ($P_Q > 0.75$)、1,033,486 ($P_Q > 0.90$) 和 798,928($P_Q > 0.95$)。如果只考虑完备性，我们可以结合二元与多元的预测结果，而如果考虑高可靠性，我们可以采用各种不同分类器相同的分类结果，这种情况下，类星体候选体的数量共为 1,262,964, 其中置信度大于 75%、90%、95% 的天体分别有 694,260 ($P_Q > 0.75$)、375,591 ($P_Q > 0.90$)、235,713 ($P_Q > 0.95$)。根据 BASS DR3 的分布天区，预测的类星体候选体的数量与通过类星体的光度函数计算的数量基本一致 (Palanque-Delabrouille et al., 2016)。

我们对具有红外信息的样本作了进一步的分析。图 3-6 显示了使用二元分类器时获得的不同概率的类星体候选体的数量；图 3-7 显示了使用不同分类器时获得的概率大于 95% 的类星体候选体的数量与 r 星等的关系；图 3-8 显示了使用二元和多元分类器时获得的概率都高于 95% 的类星体候选体在银河坐标系空间的分布。当 $r > 23$ 时，类星体候选体的数量明显减少。正如图 3-8 所示，大多数类星体候选体分布在中纬度和高纬度，沿银道面没有类星体候选体的堆积，符合类星体的分布预期。

SIMBAD（Set of Identifications, Measurements, and Bibliography for Astronomical Data）[1] 是由法国斯特拉斯堡天文数据中心 (CDS）负责运行维护的一个天文数据库，其中主要提供有关科学文章中研究过的天体的信息，包括天体位置、类型、星等、红移等经确认过的天体信息。截止到 2020 年 6 月，SIMBAD 数据库包含了大约 58,000,000 颗恒星、5,500,000 个非恒星对象（如星系、星云、星团、超新星等），其中收录的天体数据都是比较准确。我们将获得的可能性超过 95% 的类星体候选体与 SIMBAD 数据库进行交叉，交叉半径为 1 角秒，交叉结果得到 184,376 个天体。其中 SIMBAD 数据库中明确为类星体的有 175,292 个，达 95%。

[1] http://simbad.cds.unistra.fr/guide/simbad.htx

表 3-9 BASS DR3 天体分类结果星表示列。

Table 3-9 The sample of predicted results of BASS DR3 sources.

id	ra	dec	Class_*b*	P_{b1}	P_{b2}	Class_*m*	P_m	Class_*bi*	P_{bi1}	P_{bi2}	Class_*mi*	P_{mi}
95373011289	135.04541765729505	84.39853006107745	0	0.992	0.993	0	0.981	0	0.997	1.000	0	0.999
95373012270	137.06618290826282	84.45571948189149	0	0.999	0.965	0	0.994	0	0.999	0.978	0	0.993
95373014038	135.24278336054908	84.55902133199768	0	0.962	0.999	0	0.963	0	1.000	1.000	0	0.998
95374009915	145.63427167809962	84.3344909394295	0	0.976	0.999	0	0.980	0	0.956	0.999	0	0.993
95374010728	145.64322900143938	84.37462462607955	0	0.997	0.993	0	0.956	0	1.000	1.000	0	0.993
95375006879	146.35493428612986	84.18153150948889	0	0.999	0.996	0	0.983	0	1.000	1.000	0	1.000
95375008769	146.7341471886791	84.28327327388713	0	0.963	0.980	0	0.970	0	0.999	1.000	0	0.998
95375010415	146.783645308867	84.37315314289253	0	0.999	0.936	0	0.943	0	1.000	1.000	0	0.999
95371004745	123.71207464080723	84.1579130418392	1	1.000	0.998	1	0.999	1	1.000	1.000	1	1.000
95371004752	123.76995258941326	84.15844883612738	1	0.989	0.996	1	0.961	1	0.998	0.999	1	0.997
95371005233	123.80162434791758	84.18398863492929	1	0.999	0.999	1	0.998	1	1.000	1.000	1	1.000
95371005325	123.79929150615568	84.18879820615973	1	0.996	0.938	1	0.963	1	0.999	0.999	1	0.999
95371005366	123.70596853052086	84.19100324818724	1	0.970	0.997	1	0.970	1	0.972	0.997	1	0.998
95371005498	123.91422908428144	84.19767799033559	2	0.965		2	0.952	2	0.946		2	0.968
95371005594	123.61974576953872	84.2038370079812	2	0.960		2	0.974	2	0.991		2	0.991
95371005759	123.91886392365679	84.21438506899153	2	0.995		2	0.995	2	0.951		2	0.990
95371005789	123.15220933495931	84.21426532956772	2	0.986		2	0.978	2	0.995		2	0.974
95371005812	123.77907625197672	84.21706698284555	2	0.995		2	0.991	2	0.996		2	0.983
95371005951	123.84673982200142	84.22579958818966	2	0.973		2	0.988	2	0.998		2	0.988
95371005960	123.77794020713335	84.22598315903795	2	0.912		2	0.970	2	0.994		2	0.991

[a] Class_b 表示采用分类器 I 及分类器 II 得到的最终分类结果标签, P_{b1} 和 P_{b2} 表示分类器 I 和 II 得到分类结果的概率。

[b] Class_m 表示采用分类器 III 得到的分类结果标签, P_m 为分类器 III 的概率。

[c] Class_bi 表示采用分类器 IV 及分类器 V 得到的最终分类结果标签, P_{bi1} 和 P_{bi2} 分别对应为分类器 IV 和 V 的概率。

[d] Class_mi 表示采用分类器 VI 得到的最终分类结果标签, P_{mi} 为该分类器的概率。

[e] 分类结果标签定义: 0 表示类星体, 1 表示恒星，2 表示星系。

[f] 对于星系, P_{b2} 或 P_{bi2} 的值为空。

表 3-10 不同分类器及不同置信度区间下恒星、星系及类星体的数量。

Table 3-10 Star, galaxy and quasar candidates by different classifiers with different information.

	光学波段			光学与红外波段		
	二元分类器	多元分类器	分类结果一致	二元分类器	多元分类器	分类结果一致
$P_S > 0.75$	21 175 837	20 550 441	19 829 533	12 938 789	12 913 835	12 785 232
$P_S > 0.90$	18 570 067	17 759 188	17 043 293	12 697 599	12 711 753	12 561 500
$P_S > 0.95$	16 706 080	15 727 548	15 022 399	12 519 421	12 560 931	12 375 838
$P_G > 0.75$	54 360 301	56 235 525	49 483 839	25 929 229	26 490 981	25 068 898
$P_G > 0.90$	41 243 840	41 713 806	35 544 793	23 417 967	23 899 117	21 890 547
$P_G > 0.95$	32 053 408	31 025 081	25 949 348	20 888 403	21 088 855	18 606 073
$P_Q > 0.75$	11 575 871	9 459 579	7 095 580	2 078 981	1 777 386	1 500 099
$P_Q > 0.90$	5 271 397	4 052 081	2 775 970	1 419 024	1 221 362	1 033 486
$P_Q > 0.95$	2 674 704	1 874 389	1 166 517	1 088 976	943 486	798 928

[a] P_S 表示被预测为恒星的概率，对于只有光学信息的天体，如果 $P_S > 0.95$，则表明 P_{b1} 和 P_{b2} 同时大于 95%。而对于同时具有光学与红外信息的样本，则说明 P_{bi1} 和 P_{bi2} 同时大于 95%;

[b] P_G 表示被预测为星系的概率，对于只有光学信息的天体，如果 $P_G > 0.95$，则表明 P_{b1} 大于 95%。而对于同时具有光学与红外信息的天体，如果 $P_G > 0.95$，则说明 P_{bi1} 大于 95%;

[c] P_Q 表示被预测为类星体的概率, 对于只有光学信息的天体，如果 $P_Q > 0.95$，则说明 P_{b1} 和 P_{b2} 都大于 95%，而对于同时具有光学与红外信息的天体，如果 $P_Q > 0.95$，则表明 P_{bi1} 和 P_{bi2} 都大于 95%。

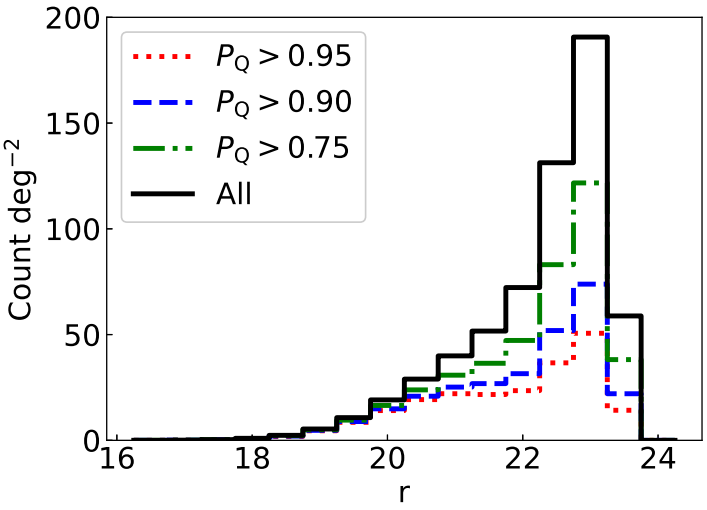


图 3-6 基于可见光和红外信息，采用二元分类器得到的具有不同概率的类星体候选体随 *r* 星等的分布密度。

Figure 3-6 The number of quasar candidates by binary classifier with optical and infrared information as function of *r* magnitude for different probabilities ($P_Q > 0.95$ (red dotted line), $P_Q > 0.90$ (blue dash line), $P_Q > 0.75$ (green dotted dash line, all (black line)).

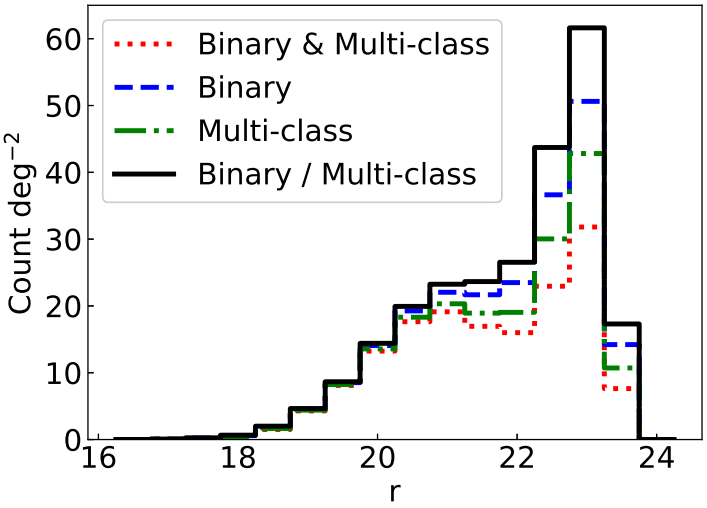


图 3-7 不同分类器得到的概率大于 95% 的类星体候选体随 *r* 星等的分布密度。

Figure 3-7 The number of quasar candidates as function of *r* magnitude with larger than 95 per cent probability by binary&multiclass (red dotted line), binary (blue dash line), multiclass (green dotted dash line) and binary/multiclass (black line) classifiers based on optical and infrared information.

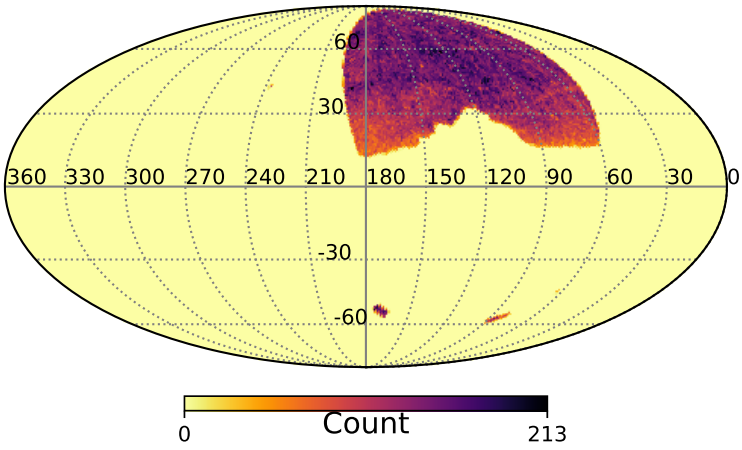


图 3-8 基于可见光和红外信息，采用二元和多元分类器得到的概率都大于 95% 的类星体候选体在银河坐标系空间的分布。

Figure 3-8 The galactic location of quasar candidates with larger than 95 per cent probability by binary&multiclass classifiers based on optical and infrared information.

如果加上 AGN 等具有特殊发射线的天体，则达到了 97%。验证结果进一步表明，研究中构建的分类器具有较强的泛化能力，分类结果是可靠的。

3.6　本章小结

本章我们采用来自 SDSS DR16、DR16Q 及 LAMOST DR5 的光谱数据与 BASS DR3 的 g、r、z，ALLWISE 的 $W1$、$W2$ 组成的光学与红外测光数据一起构建已知星系、恒星和类星体样本，利用 XGBoost 学习算法进行了二元与多元的分类方法研究。我们分别进行了分类模型的输入特征与超参数的最优化，创建了不同分类目标和输入特征的最优模型。在综合考虑光学与红外特征时，10 折交叉验证的平均准确率均超过了 95%。然后我们分别采用二元及多元两种分类策略对 BASS DR3 的数据进行了分类。对于包含光学与红外信息的数据，在采用二元与多元分类两种策略下结果都是类星体且分类概率大于 95% 的类星体数据为 798,928。我们再将这些数据与 SIMBAD 数据库进行交叉验证，交叉结果中 95% 的天体分类与预测类型相一致。我们将分类结果分别提交给了云南 2.4 米望远镜及 LAMOST 望远镜，期望后续能够有更多的类星体得到证认。

第 4 章　类星体的测光红移

红移是宇宙天体的重要特征，尤其是河外天体。红移的测量，直接反映了天体的距离，对于研究天体的空间位置、形成与演化、光度函数及宇宙大尺度结构的研究等均具有重要的意义。测光红移则是根据大型巡天中获取的天体多波段测光数据来进行红移的估计。近年来，随着许多大型巡天望远镜的投入运行，我们已经积累了大量的宇宙天体的多波段信息，如果能够将这些天体的红移准确测量出来，将可以绘制出宇宙的三维空间分布。SDSS、LAMOST 等大规模光谱巡天数据为我们提供了大量红移在 0 到 6 之间的类星体的准确红移数据，为测光红移算法的研究提供了很好的训练样本。测光红移估计方法大致可以分为两类：经验方法和模板匹配方法。经验方法通常是指在已有光谱红移的基础上，经过训练，构建红移与多波段测光数据之间的关系模型，再将构建的关系模型为未知的天体预测红移。而模版匹配方法通常使用实测光谱或模拟光谱作为模版库，然后与测光巡天望远镜的滤光片的响应曲线卷积，从而创建出测光红移估计的模板，即红移、多波段流量（星等）、颜色及误差之间的关系网格。这两类方法各有优缺点，在测光红移研究中都有着广泛的应用。本研究将基于机器学习算法，构建最优的回归器并完成对上一节的类星体候选体的红移预测。

4.1　数据

SDSS DR16Q 发布了近 75 万颗类星体，是迄今为止最大的类星体样本库。LAMOST DR5 发布了大约 5 万颗类星体，这些类星体都是经过光谱证认的，提供了准确的光谱红移。为了简化，我们从前一章预测的 BASS DR3 的结果星表中选择所有的类星体侯选体，即 $Class\_b, Class\_bi, Class\_m$ 及 $Class\_mi$ 中有一个值为 0，这些结果星表中已经包含了光学（BASS DR3 的 g、r、z 三个波段）及红外波段（ALLWISE 的 $W1$、$W2$ 波段）的数据，数据总数为 26,200,778。然后我们将 SDSS DR16Q 及 LAMOST DR5 中的类星体分别与这些选择的数据交叉，交叉半径为 2 角秒。交叉结果分别记为 BS_W 和 BL_W。然后，根据是否具有有效的红外信息，即 $W1mag$ 和 $W2mag$ 的值是否为 $NULL$ 或者空格，我们把样本 BS_W 分成 BSO 和 BSW 两个样本。同样的，BL_W 也分成 BLO 和 BLW 两个样本。我们把样本 BS_W 和 BSW 用作训练样本和测试样本，而把样本 BL_W 和 BLW 用作外部测试, 以进一步检测回归器的性能及泛化能力。图4-1为训练样本的红移分布；图4-2为测试样本的红移分布。从图4-1及图4-2可以发现，训练样本与测试样本的红移范围是一致的。

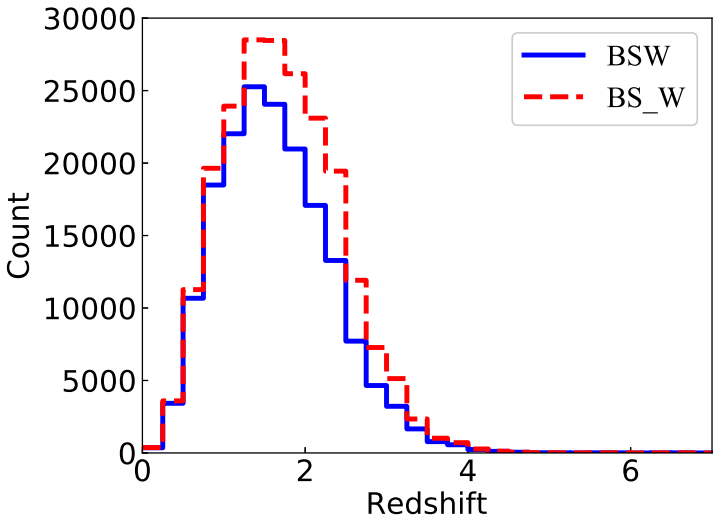


图 4-1 样本 BSW 和 BS_W 的光谱红移分布，蓝色表示 BSW 样本、红色表示 BS_W 样本。
Figure 4-1 The distribution of spectroscopic redshifts for known samples BSW (blue line) & BS_W (red dash line).

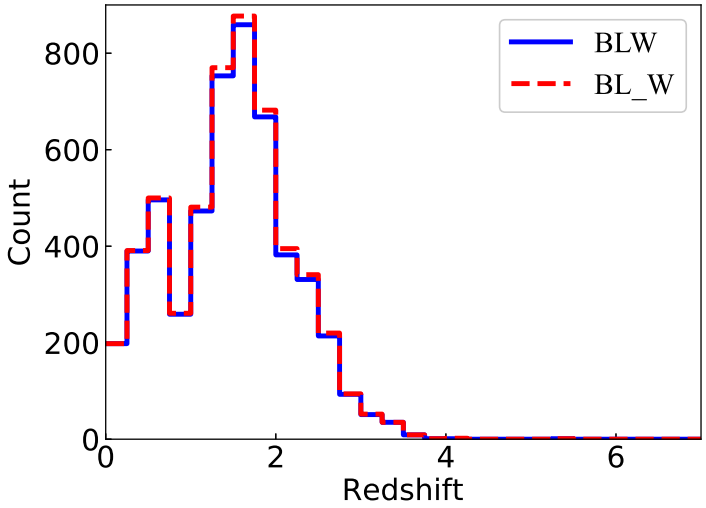


图 4-2 样本 BLW 和 BL_W 的光谱红移分布，蓝色表示 BLW 样本、红色表示 BL_W 样本。
Figure 4-2 The distribution of spectroscopic redshifts for known samples BLW (blue line) & BL_W (red dash line).

4.2 回归算法性能评价指标

本研究主要采用 CatBoost、XGBoost 和随机森林进行回归模型的构建，然后选择最优的模型用于对类星体候选体进行测光红移估计。关于方法的详细介绍可参考第二章。为了比较不同方法在回归分析的性能，我们采用了多个不同的指标来进行综合评价。首先，我们定义残差 (Δz) 为光谱红移与测光红移之差，即

$\Delta z = z_{spec} - z_{photo}$。平均绝对误差 (MAE)、均方差 (MSE) 定义如下：

$$MAE = \frac{1}{n}\sum_{i=0}^{n-1}|z_i - \hat{z}_i| \tag{4-1}$$

$$MSE = \frac{1}{n}\sum_{i=0}^{n-1}(z_i - \hat{z}_i)^2 \tag{4-2}$$

公式中 z_i 表示光谱红移, $\hat{z}_i$ 表示预测的测光红移，n 表示样本数量。

在许多机器学习的测光红移工作中，规范化残差也经常被实际采用，它的定义如下：

$$\Delta z(\mathrm{norm}) = \frac{z_{spec} - z_{phot}}{1 + z_{spec}} \tag{4-3}$$

规范化残差的绝对值小于某一给定值的样本数量占总样本量的比例，也常用作评价回归算法优劣的指标 (Schindler et al., 2007)。例如，当给定值为 0.3 时，其定义如下：

$$\delta_{0.3} = \frac{N_{|\Delta z(\mathrm{norm})|<0.3}}{N_{total}} \tag{4-4}$$

其它使用指标还有相关系数（R^2）、偏差（$Bias$）、标准偏差（$\sigma_{\Delta z}$）、规范化的中值绝对偏差 (σ_{NMAD}) 和离群率 (O) (Henghes et al., 2021; Curran et al., 2021), 分别定义如下：

$$R^2 = 1 - \frac{\sum_{i=0}^{n-1}(z_i - \hat{z}_i)^2}{\sum_{i=0}^{n-1}(z_i - \overline{z})^2} \tag{4-5}$$

$$\mathrm{Bias} = < z_{spec} - z_{phot} > \tag{4-6}$$

$$\sigma_{\Delta z} = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(\Delta z)^2} \tag{4-7}$$

$$\sigma_{NMAD} = 1.48 \times \mathrm{median}\left|\frac{z_{spec} - z_{phot}}{1 + z_{spec}}\right| \tag{4-8}$$

$$Outlier\ fraction(O) = \frac{N_{|\Delta z(\mathrm{norm})|>0.15}}{N_{total}} \tag{4-9}$$

4.3　最优特征选择

不论分类还是回归，寻找最优特征都是必不可少的过程。输入特征不仅是影响机器学习算法性能的关键因素，而且通过去掉不重要的特征，降低数据维度，从而提升训练速度，同时，也有助于改善算法的性能。由于样本中存在很大部分没有红外对应体的数据，为了方便描述，我们将所有特征分为光学特征及红

外特征两类，光学特征包括 $\Delta g, \Delta r, \Delta z$, $g, r, z, g-r, r-z, g-z$, 红外特征包括 $W1, W2, g-W1, r-W1, z-W1, g-W2, r-W2, z-W2, W1-W2$。对于样本 BSW 和 BS_W, 它们包括了所有的光学与红外特征，而样本 BSO 中则只包含光学特征。样本 BSW 与 BS_W 的区别在于 BS_W 除 BSW 样本外，其它数据的红外信息为空。因此，样本 BS_W 和 BL_W 包含了缺值数据。算法 XGBoost 和 CatBoost 支持对缺值数据的训练，而当使用随机森林时，将缺值设为 0。

CatBoost、XGBoost 和随机森林三种方法都具备特征重要性的评价能力，我们选择每个特征的总增益作为特征重要性的评价标准。图4-3和图4-4分别展示了三种学习方法在样本 BSW 和 BS_W 上得到的特征重要性的排列顺序。图4-5显示的是三种学习方法在样本 BS_W 上只采用光学特征时的特征重要性排列顺序。

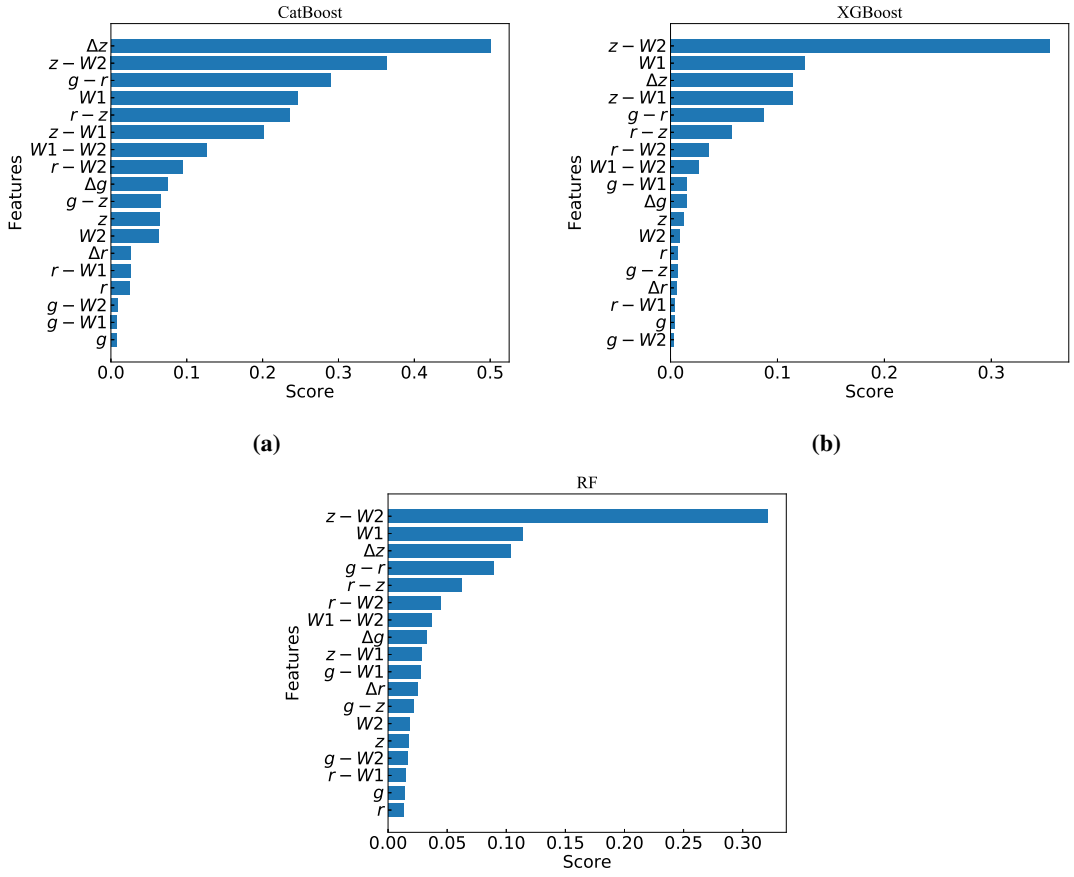


图 4-3　CatBoost、XGBoost 和随机森林方法在数据样本 BSW 上给出的特征重要性排序。

Figure 4-3　The feature importance rank for Sample BSW by CatBoost, XGBoost and Random Forest.

从图4-3、图4-4和图4-5的特征重要性排列表明，特征重要性与样本数据、算法紧密相关。但也同样具有一些共性的地方。例如当采用光学与红外特征相结合时，三种方法排序最靠前的 4 个特征中有三个特征（$z-W2, \Delta z, W1$）是一样的，而当只采用光学特征时，三种方法给出的特征重要性排序中靠前的 4 个特

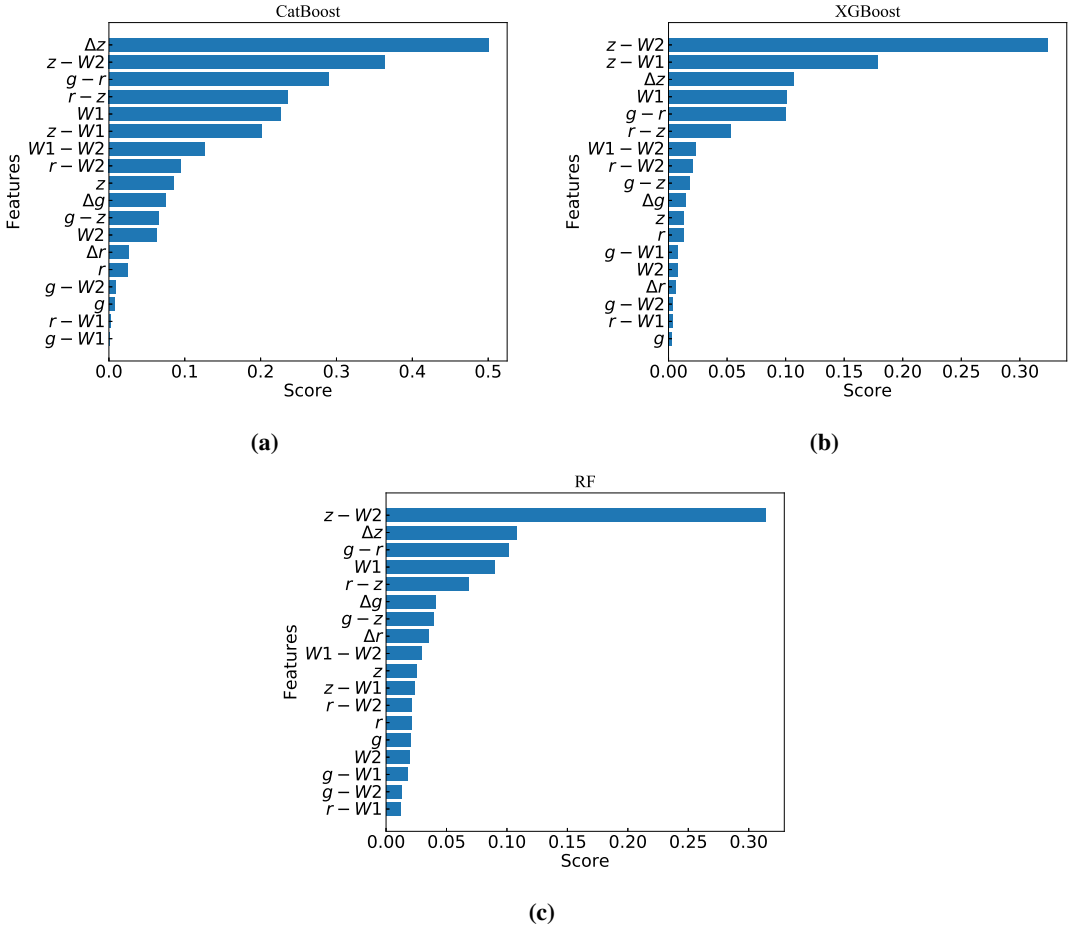


图 4-4　CatBoost、XGBoost 和随机森林方法在数据样本 BS_W 上给出的特征重要性排序。

Figure 4-4　The feature importance rank for Sample BS_W by CatBoost, XGBoost and Random Forest.

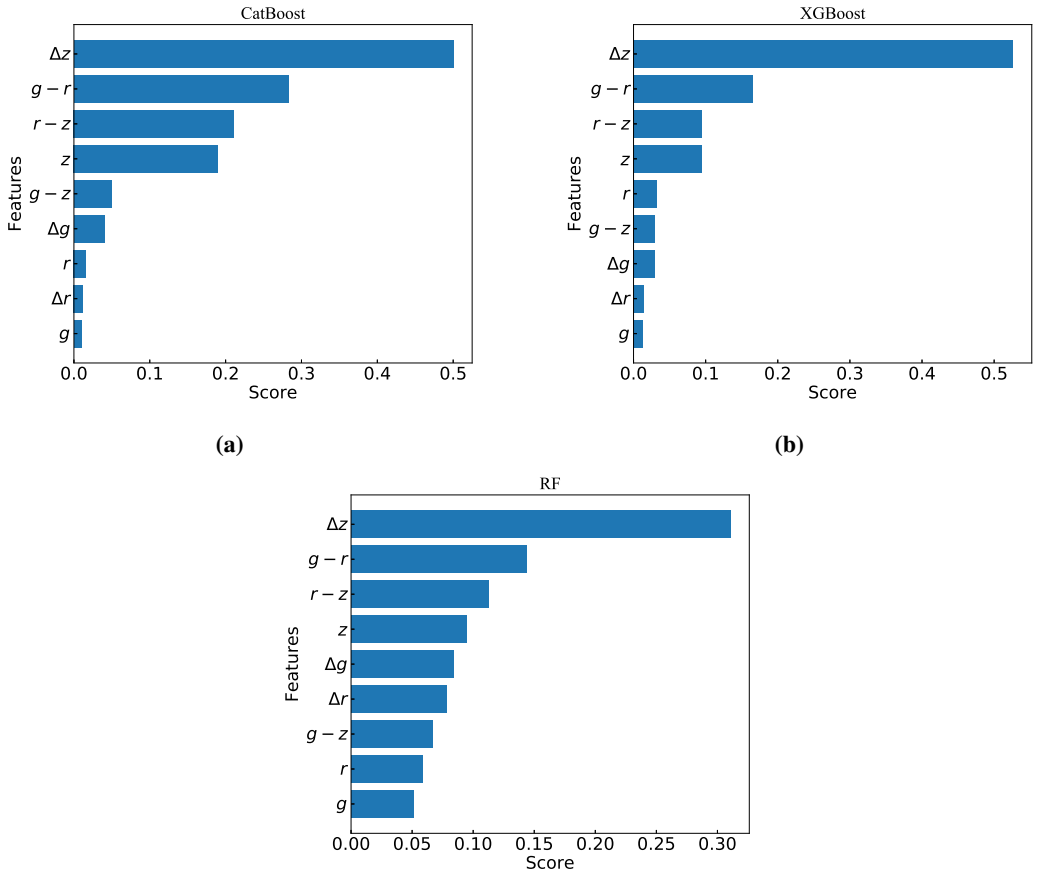


图 4-5　CatBoost、XGBoost 和随机森林方法在数据样本 BS_W 上给出的只有光学特征时的特征重要性排序。

Figure 4-5　The only optical feature importance rank for Sample BS_W by CatBoost, XGBoost and Random Forest.

征（Δz, $g-r$, $r-z$, z）也都是一样的。这表明特征重要性的排序虽然跟算法有关，但数据自身的特点更为重要。

特征重要性基本确定了输入特征的顺序，但并不是所有特征全部使用时，回归预测的性能就一定会好。为了寻找最优的输入特征，我们设计了一个最优特征寻找算法。我们先选择最重要的 4 个特征作为初始输入特征，并以缺省的超参数进行模型的训练，采用五折交叉的模型验证方法，记录五折平均的性能，采用的性能指标包括 MSE、MAE、bias、σ_{NMAD}、$\sigma_{\Delta \mathrm{z}}$、$R^2$、$\delta_{0.3}$、$Outlier\ fraction\,(O)$ 和运行时间。然后我们依次向输入特征中增加一个未使用的最重要特征，进行同样的训练。当完成所有训练后，我们比较每次训练的回归器性能，以 MSE 指标为主，其它为辅。表4-1列出了每种方法在每个样本上的最好性能时的详细指标；图4-6展示了每次训练的 MSE 值变化图。我们最终选择 MSE 值最小的一组输入特征作为最优。对于样本 BSW，使用 XGBoost 算法时，最优输入特征为 $z-W2$, $W1$, Δz, $z-W1$, $g-r$, $r-z$, $r-W2$, $W1-W2$, $g-W1$, Δg, z (11 features); 使用 CatBoost 算法时，最优输入特征为 Δz, $z-W2$, $g-r$, $W1$, $r-z$, $z-W1$, $W1-W2$, $r-W2$, Δg, $g-z$, z, $W2$, Δr (13 features)；使用随机森林时，最优输入特征为 $z-W2$, $W1$, Δz, $g-r$, $r-z$, $r-W2$, $W1-W2$, Δg, $z-W1$, $g-W1$, Δr, $g-z$, $W2$, z, $g-W2$ (15 features)。对于样本 BS_W, 使用 XGBoost、CatBoost 和随机森林时的最优输入特征分别为 $z-W2$, $z-W1$, Δz, $W1$, $g-r$, $r-z$, $W1-W2$, $r-W2$, $g-z$, Δg, z, r, $g-W1$, $W2$, Δr, $g-W2$ (16 features); Δz, $z-W2$, $g-r$, $r-z$, $W1$, $z-W1$, $W1-W2$, $r-W2$, z, Δg, $g-z$, $W2$, Δr, r, $g-W2$ (15 features)；$z-W2$, Δz, $g-r$, $W1$, $r-z$, Δg, $g-z$, Δr, $W1-W2$, z, $z-W1$, $r-W2$ (12 features) 。而如果只考虑光学特征，保留所有光学特征并按重要性排序时，三种方法都分别达到最优的预测性能。综合比较各种评价指标，在使用缺省超参数的情况下，CatBoost 的性能要优于 XGBoost 和随机森林。

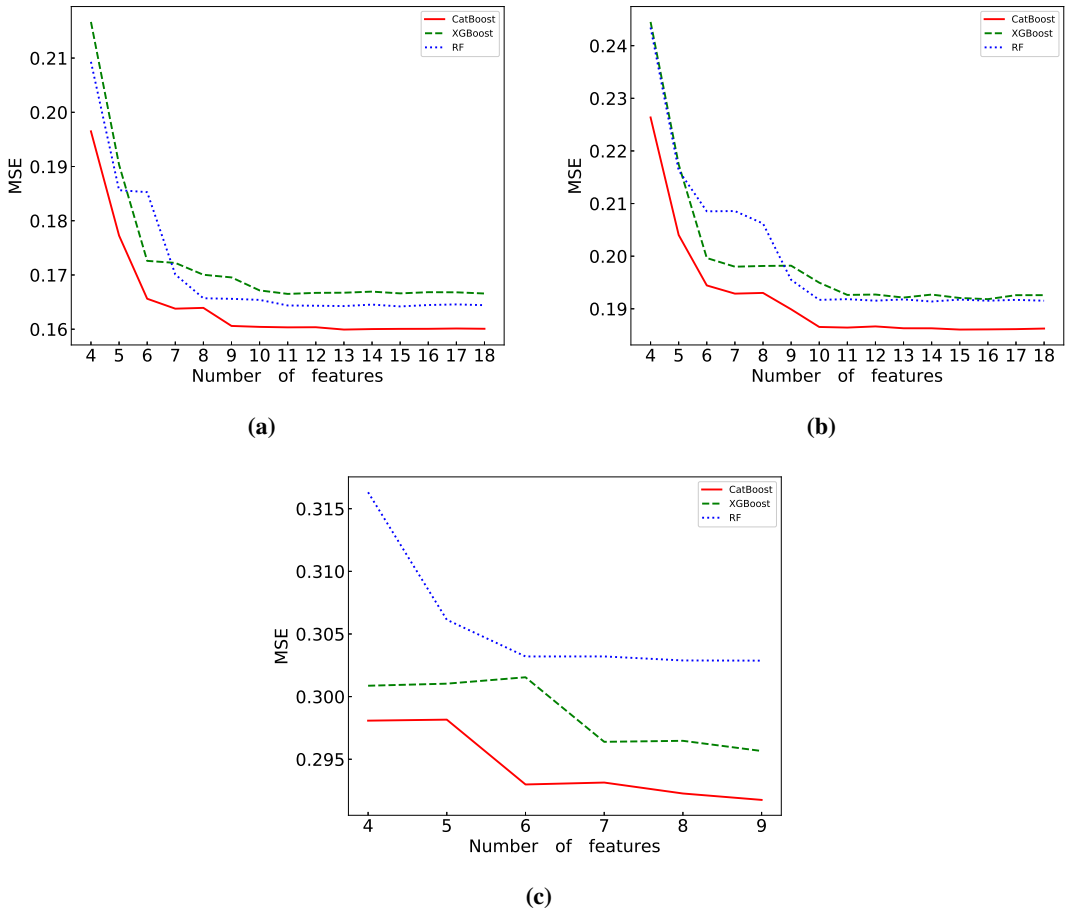


图 4-6 不同样本在不同学习算法及输入特征时的性能比较。(a) 为 BSW 样本（光学与红外特征）的实验结果；(b) 为 BS_W 样本（光学与红外特征）的实验结果；(c) 为 BS_W 样本（只采用光学特征）的实验结果。

Figure 4-6 The performance of different methods with different input patterns for different samples. Panel (a): for the sample BSW with optical and infrared information; panel (b): for the sample BS_W with optical and infrared information; panel (c): for the sample BS_W only with optical information.

表 4-1 XGBoost、CatBoost 及随机森林在缺省模型参数时的最优性能。

Table 4-1 The performance of photometric redshift estimation with optimal input features for each method with default model parameters.

样本	输入特征	方法	MSE	MAE	Bias	$\sigma_{\rm NMAD}$	$\sigma_{\Delta z}$	R^2	$\delta_{0.3}$(%)	O(%)	Time(s)
BSW	Pattern I	XGBoost	0.1666	0.2765	4.0×10^{-5}	0.1070	0.4082	0.6493	93.39	22.04	3
BSW	Pattern II	CatBoost	0.1600	0.2702	6.0×10^{-5}	0.1036	0.4000	0.6632	93.61	21.32	10
BSW	Pattern III	RF	0.1645	0.2700	2.2×10^{-3}	0.1008	0.4056	0.6537	93.35	21.01	442
BS_W	Pattern IV	XGBoost	0.1925	0.3064	-7.6×10^{-6}	0.1171	0.4387	0.6071	92.37	25.22	6
BS_W	Pattern V	CatBoost	0.1867	0.3012	-4.8×10^{-6}	0.1146	0.4321	0.6189	92.56	24.48	12
BS_W	Pattern VI	RF	0.1924	0.3009	-2.0×10^{-5}	0.1107	0.4387	0.6074	92.32	24.03	447
BS_W	Pattern VII	XGBoost	0.2959	0.4009	-1.0×10^{-4}	0.1657	0.5439	0.4008	87.27	37.40	4
BS_W	Pattern VII	CatBoost	0.2917	0.3988	-9.0×10^{-5}	0.1651	0.5403	0.4087	87.39	37.74	12
BS_W	Pattern VII	RF	0.3032	0.4045	8.7×10^{-3}	0.1648	0.5507	0.3858	86.52	37.93	200

[a] Pattern I 代表 $z-W2, W1, \Delta z, z-W1, g-r, r-z, r-W2, W1-W2, g-W1, \Delta g, z$ (11 features).

[b] Pattern II 代表 $\Delta z, z-W2, g-r, W1, r-z, z-W1, W1-W2, r-W2, \Delta g, g-z, z, W2, \Delta r$ (13 features).

[c] Pattern III 代表 $z-W2, W1, \Delta z, g-r, r-z, r-W2, W1-W2, \Delta g, z-W1, g-W1, \Delta r, g-z, W2, z, g-W2$ (15 features).

[d] Pattern IV 代表 $z-W2, z-W1, \Delta z, W1, g-r, r-z, W1-W2, r-W2, g-z, \Delta g, z, r, g-W1, W2, \Delta r, g-W2$ (16 features).

[e] Pattern V 代表 $\Delta z, z-W2, g-r, r-z, W1, z-W1, W1-W2, r-W2, z, \Delta g, g-z, W2, \Delta r, r, g-W2$ (15 features).

[f] Pattern VI 代表 $z-W2, \Delta z, g-r, W1, r-z, \Delta g, g-z, \Delta r, W1-W2, z, z-W1, r-W2$ (12 features).

[g] Pattern VII 代表 $\Delta z, \Delta g, \Delta r, g-r, g-z, r-z, g, r, z$ (9 features).

[h] Pattern I, II, III, IV, V, VI, VII 在后续表述中表示相同的意义。

4.4 一步回归模型构建

一步回归模型是指对整个样本进行训练，只构建一个回归模型来对所有样本数据进行红移估计。模型参数优化是模型构建的主要任务，也是一项非常复杂的工作。为了减少计算规模，我们选择了一些主要的超参数进行优化，包括树的深度及数量。我们采用网格搜索（grid search）及 5-折交叉验证的方法进行训练与评价，采用的评价指标包括 MSE、MAE、$Bias$、σ_{NMAD}、$\sigma_{\Delta z}$、R^2、$\delta_{0.3}$、O 及运行时间。表 4-2列出了最优参数时的详细性能指标。通过比较表中详细的性能数据，CatBoost 方法显示出它的优越性。

然后，我们采用 8:2 的比例将样本 BS_W 和 BSW 分为训练集与测试集，采用表4-2中各种方法获得的最优超参数分别进行训练与测试。图4-7和图4-8展示了测试样本在光谱红移与测光红移二维空间上的散点分布图及 Δz(norm) 分布。

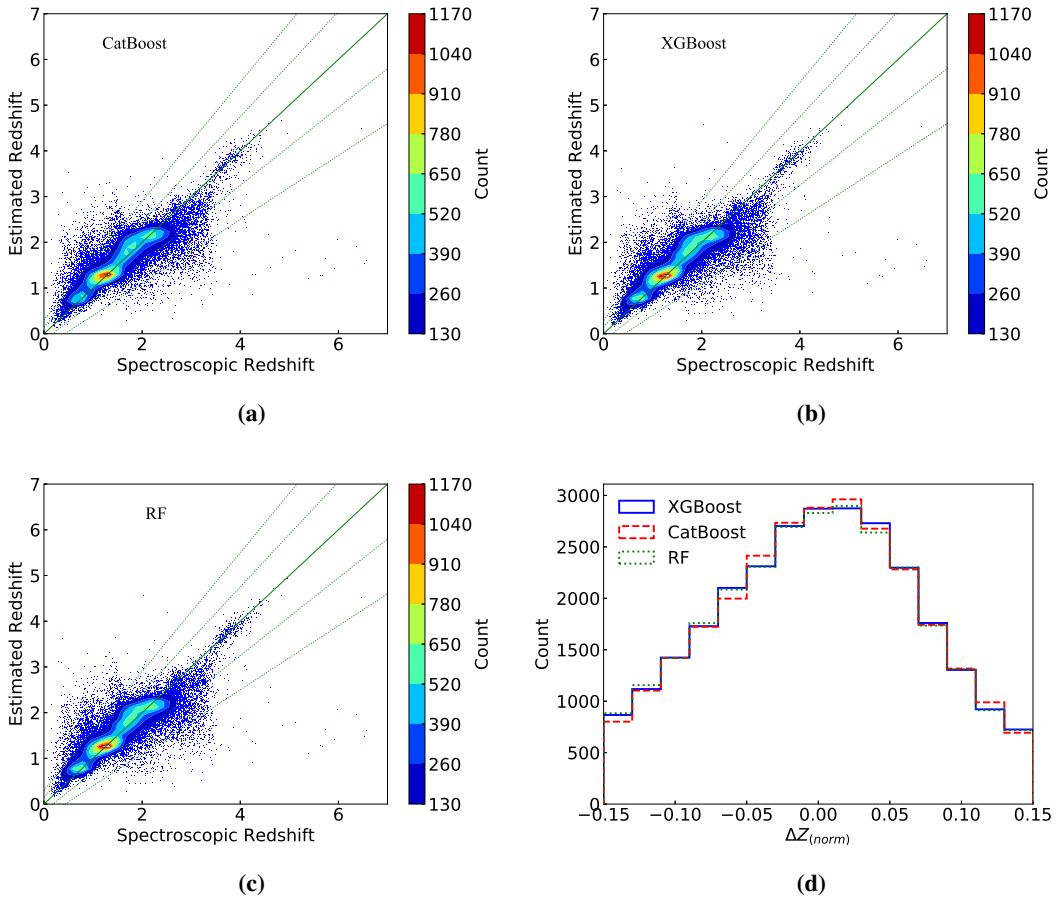


图 4-7 CatBoost、XGBoost 及随机森林三种方法基于样本 BSW 的测光红移与光谱红移的散点分布图及 Δz(norm) 分布图。

Figure 4-7 The performance of photometric redshift estimation with CatBoost, XGBoost and RF for sample BSW.

最后，我们分别使用最优模型参数对整个样本进行训练，再使用样本 BL_W 和 BLW 作为外部测试样本进行回归器性能的验证。表4-3列出了 CatBoost 在不

表 4-2 一步模型下不同分类器的 5 折交叉最优性能。

Table 4-2 The performance of photometric redshift estimation with the best features and optimal model parameters.

训练样本	输入特征	学习方法	模型参数	MSE	MAE	Bias	σ_{NMAD}	$\sigma_{\Delta z}$	R^2	$\delta_{0.3}$(%)	O(%)	训练时间 (s)
BSW	Pattern I	XGBoost	max_depth=11 n_estimators=1000	0.1617	0.2669	-0.0009	0.1001	0.4022	0.6595	93.61	20.58	110
BSW	Pattern II	CatBoost	depth=12 iterations=4000	0.1576	0.2649	-0.0006	0.0999	0.3971	0.6680	93.78	20.45	421
BSW	Pattern III	RF	max_depth=15 n_estimators=500	0.1626	0.2686	0.0001	0.1007	0.4032	0.6578	93.50	20.83	1336
BS_W	Pattern IV	XGBoost	max_depth=10 n_estimators=1200	0.1865	0.2969	0.0001	0.1106	0.4318	0.6195	92.62	23.72	325
BS_W	Pattern V	CatBoost	depth=12 iterations=4000	0.1848	0.2960	0.0001	0.1105	0.4299	0.6228	92.69	23.55	468
BS_W	Pattern VI	RF	max_depth=15 n_estimators=500	0.1893	0.2990	0.0009	0.1111	0.4351	0.6136	92.44	23.82	1431
BS_W	Pattern VII	XGBoost	max_depth=10 n_estimators=1000	0.2930	0.3967	-0.0001	0.1626	0.5413	0.4066	87.43	37.17	110
BS_W	Pattern VII	CatBoost	depth=12 iterations=2000	0.2906	0.3970	-0.0001	0.1637	0.5393	0.4110	87.47	37.38	254
BS_W	Pattern VII	RF	max_depth=15 n_estimators=500	0.2944	0.3975	0.0003	0.1623	0.5425	0.4038	87.36	37.07	1124

同的输入模型、不同的训练样本和测试样本下的性能。由表4-3可知，当训练样本为 BSW 时，CatBoost 自验证的最优性能为 $MSE = 0.1059$, $MAE = 0.2223$, $Bias = -1.6 \times 10^{-5}$, $\sigma_{\mathrm{NMAD}} = 0.0872$, $\sigma_{\Delta z} = 0.3254$, $R^2 = 0.7780$, $\delta_{0.3} = 96.01\%$ 和 $O = 15.79\%$，而采用 BLW 样本进行外部验证时的性能为 $MSE = 0.1239$, $MAE = 0.2134$, $Bias = 0.0265$, $\sigma_{\mathrm{NMAD}} = 0.0797$, $\sigma_{\Delta z} = 0.3520$, $R^2 = 0.7585$, $\delta_{0.3} = 94.50\%$ 和 $O = 15.11\%$。当训练样本为 BS_W、输入特征为 Pattern V 时的模型自验证及外部验证的性能都要优于输入特征为 Pattern VII 的模型性能。因此，我们可以采用此模型来预测所有只含有光学特征的类星体红移，对应红外波段的数据作缺值处理。

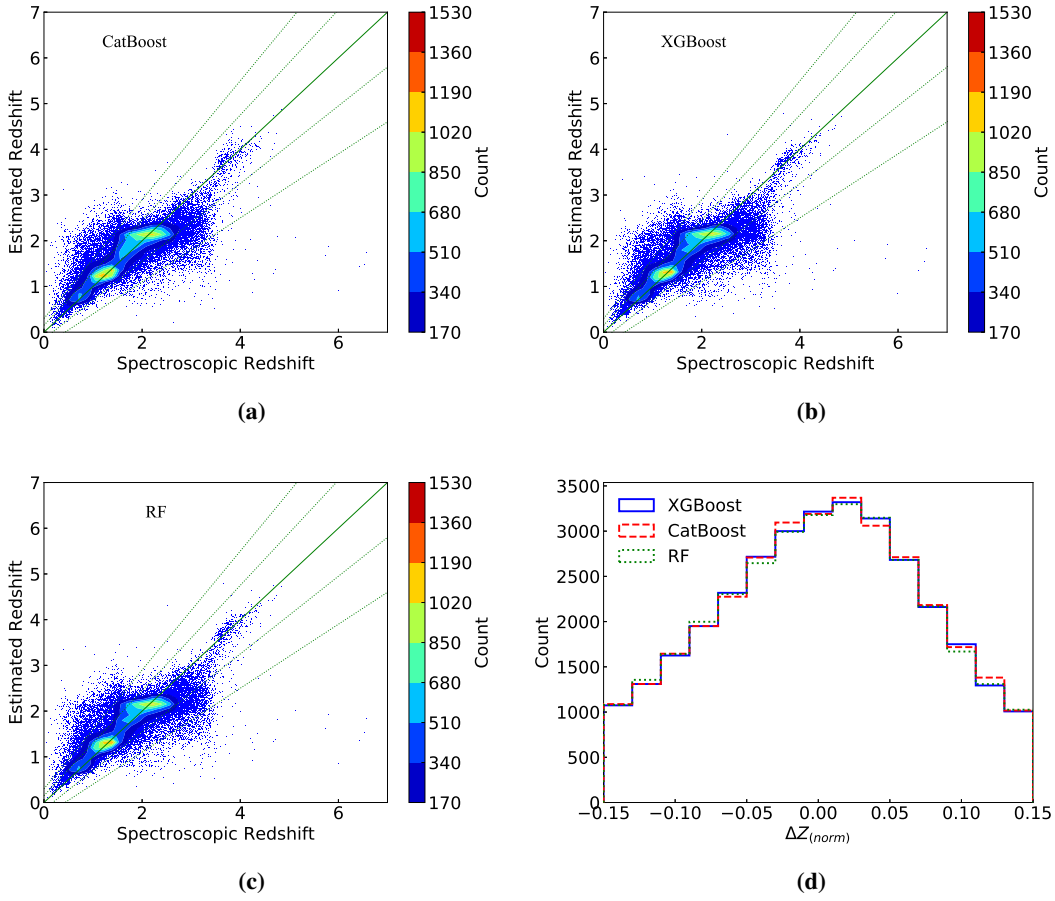


图 4-8 CatBoost、XGBoost 及随机森林三种方法基于样本 BS_W 的测光红移与光谱红移的散点分布图及 Δz(norm) 分布图。

Figure 4-8 The performance of photometric redshift estimation with CatBoost, XGBoost and RF for sample BS_W.

表 4-3 CatBoost 算法进行红移估计的最优性能。

Table 4-3 The performance of CatBoost for photometric redshift estimation.

训练样本	输入特征	测试样本	MSE	MAE	Bias	σ_{NMAD}	$\sigma_{\Delta z}$	R^2	$\delta_{0.3}(\%)$	$O(\%)$
BSW	Pattern II	BSW	0.1059	0.2223	-1.6×10^{-5}	0.0872	0.3254	0.7780	96.01	15.79
BSW	Pattern II	BLW	0.1239	0.2134	0.0265	0.0797	0.3520	0.7585	94.50	15.11
BS_W	Pattern V	BSW	0.1115	0.2279	-4.0×10^{-5}	0.0889	0.3340	0.7661	95.77	16.42
BS_W	Pattern V	BS_W	0.1365	0.2578	-7.0×10^{-6}	0.0981	0.3695	0.7239	94.84	19.42
BS_W	Pattern V	BLW	0.1265	0.2167	0.0268	0.0808	0.3557	0.7534	94.48	15.61
BS_W	Pattern V	BL_W	0.1277	0.2188	0.0296	0.0817	0.3574	0.7502	94.41	15.87
BS_W	Pattern V	BSO	0.2490	0.3929	1.2×10^{-4}	0.1550	0.4991	0.3884	90.64	32.93
BS_W	Pattern VII	BSO	0.3267	0.4363	-0.1853	0.1672	0.5719	0.1970	90.98	36.97
BS_W	Pattern VII	BL_W	0.2596	0.3681	0.1223	0.1639	0.5095	0.4925	85.82	37.61

4.5 两步组合模型构建

由已知样本的红移分布图 4-1 可知，样本中高红移（redshift ≥ 3.5）部分的比例远小于低红移 (redshift < 3.5) 样本的数量。例如，在样本 BS_W 的 213,359 个类星体中，只有 2,238 个高红移类星体，而在样本 BL_W 的 5,310 个类星体中只有 13 个红移大于 3.5 的类星体。因此，我们将样本 BS_W 分为高低红移两个子样本 BS_W_H 和 BS_W_L, 相应的将 BSW 分为 BSW_H 和 BSW_L。然后我们使用一步模型中的预测模型分别对高、低红移样本进行预测。表4-4列出了各子样本进行模型验证的详细性能指标。由表4-4可见，模型对高红移部分的预测性能明显要低于低红移样本。由图4-7和图4-8可见，很多高红移的类星体被预测成了低红移。样本 BS_W 中的 2,238 个高红移类星体只有 1,765 个预测在高红移区，占 78%，而在样本 BSW 中的 1,798 个高红移类星体中只有 1,492 个预测也在高红移区，占 83%。高红移样本预测性能中的 MSE 为 0.4079，远高于低红移部分的 0.1336。因此，改善与优化高红移类星体的预测性能对于优化整个红移预测模型具有重要意义。

造成高红移类星体预测性能不高的主要原因可能是由于高红移类星体样本数量不足，远低于低红移样本。这种样本的不均衡造成了回归器预测出现较大的偏差。因此，我们首先考虑增加高红移样本来解决样本的不平衡问题。我们采用 SMOTE（Synthetic Minority Oversampling Technique）(Chawla et al., 2002) 算法来进行样本的均衡化处理。SMOTE 方法是一种很受欢迎的过采样技术，它的工作方式是选择特征空间中靠近的样本在特征空间中的样本之间画一条线，然后在沿该线的点上绘制新样本。我们采用了 Python 库中的 imblearn.over_sample 中的 SMOTE 实现。我们偿试了将高红移样本的比例提高到总样本的 30%。然后再基于新的样本分别采用 CatBoost、XGBoost 和随机森林进行模型优化，得到最优参数后，再将样本进行 8:2 的比例进行训练与测试，用 CatBoost 方法得到的最优模型的测试性能为 MAE=0.1963, $\delta_{0.3}$=93.56%, 与一步模型相比，性能指标并没有得到显著改善。因为 SMOTE 算法虽然可以增加样本数量，同时避免随机采取造成的过拟合问题，但是 SMOTE 方法无法克服非平衡数据集的数据分布问题, 容易产生分布边缘化，而且，SMOTE 在进行近邻选择时, 存在一定的盲目性，所以很难保证样本的有效性。

4.5.1 第一步：类星体高低红移分类器

为了进一步改善模型性能，我们先将类星体分成高红移类星体和低红移类星体两部分（以红移 3.5 为界），再分别采用回归预测的方法进行实验，我们称为两步模型。两步模型的第一步就是构建分类器。我们在样本 BSW 和 BS_W 中增加一列 “label”，如果红移大于 3.5，则其值设置为 1，否则其值为 0。我们同时采用 CatBoost、XGBoost 和随机森林三种方法进行分类模型的训练，利用准确率、精度、召回率及 F1_score 等分类评价指标来对比分类器的性能, 分类器的性能评价指标参考第三章。模型训练采用网格搜索和 5 折交叉验证方法进行超参

表 4-4 最优 CatBoost 一步模型分别对高低红移样本进行红移估计的性能。

Table 4-4 The performance of photometric redshift estimation for high and low subsamples with the best CatBoost regressors.

训练样本	输入特征	测试样本	MSE	MAE	Bias	σ_{NMAD}	$\sigma_{\Delta z}$	R^2	$\delta_{0.3}$(%)	O(%)
BS_W	Pattern V	BS_W_H	0.4079	0.2799	-0.1943	0.0359	0.6387	-1.4640	96.15	8.04
BS_W	Pattern V	BS_W_L	0.1336	0.2575	0.0021	0.0989	0.3655	0.7019	94.83	19.34
BSW	Pattern II	BSW_H	0.4074	0.2575	-0.1741	0.0318	0.6383	-1.1900	96.27	6.7
BSW	Pattern II	BSW_L	0.1027	0.2219	0.0018	0.0879	0.3205	0.7591	96.00	15.89

数优化，三种学习方法的 5 折交叉验证的最好性能详细列在表4-5中。从该表可知，在准确率和 F1_score 指标上，CatBoost 要优于 XGBoost 和随机森林。而且，CatBoost 在高红移的召回率上是最优的，表明具有更好的完备性。为了发现更多的高红移类星体，对于高红移部分的高完备性是非常有必要的。

为了进一步验证三个分类器的性能，根据表4-5得到的最优模型参数，我们使用全样本分别进行训练与验证。表4-6列出了三个分类器的验证性能。从表 4-6 可以发现，当使用 CatBoost 算法时，基于训练样本 BS_W 得到的模型自验证时的准确率达到了 99.99%，高红移样本的 F1 值为 99.95%，低红移样本的 F1 值为 100%；而如果用样本 BSW 进行检测，准确率同样达到 99.99%, 高红移样本的 F1 值为 99.97，低红移样本的 F1 值为 100%，这比基于训练样本 BSW 自验证的性能（准确率为 99.96%, 高红移的 F1 值为 98.10%, 低红移的 F1 值为 99.98%）更高。这表明，基于样本 BS_W，采用光学与红外特征时得到的模型可以直接用于全样本的预测。相比于 XGBoost 和随机森林，CatBoost 算法只在基于 BSW 进行自验证时的准确率稍低，其它情况下的模型性能都更有优势，而且训练速度更快。因此，对于训练样本 BS_W 和输入特征为 Pattern V （光学与红外相结合的特征），我们最终将 CatBoost 算法作为第一步的核心算法。

4.5.2 第二步：分别构建高低红移样本的回归器

第二步则针对高红移、低红移子样本分别进行回归器的训练，选择最优的回归预测模型。我们同样采用网格搜索和 5 折交叉验证方法，三种学习算法得到的最优性能列在表4-7中。回归评价上，我们采用了与一步模型相一致的评价指标。将表4-7与表4-4相比，高低红移的预测性能都有了较大的改善，尤其对于高红移样本。比如 $\boldsymbol{MSE}$ 已从一步模型中的 0.4079 降到了 0.0756，$\sigma_{\rm NMAD}$ 从一步模型的 0.0359 降到了 0.0256。因此，高低红移分别进行训练，形成各自的预测模型，对于红移估计的性能具有显著的提升。而比较表 4-7 的性能数据，除 Bias、$\sigma_{\rm NMAD}$ 与 O 指标外，CatBoost 在其它指标上都要优于 XGBoost 和随机森林。而对于 XGBoost 和随机森林，虽然对高红移部分的 $\sigma_{\rm NMAD}$ 指标略有优势，但 CatBoost 也与它们的非常接近（0.0256 vs. 0.0253），而相比于随机森林，CatBoost 在训练时间上具有明显的优势。因此，我们选择 CatBoost 作为主要的回归算法来构建高低红移样本的预测模型。

4.5.3 两种模型的对比

对于两步模型，先构建分类器，将样本分类成高、低红移两个子样本后，再针对高、低红移子样本分别构建回归器。根据上述实验结果，我们采用 CatBoost 作为主要的学习算法，先分类，再分别针对两个子样本做回归。然后我们采用两步模型对已知样本进行性能验证，已知样本包括 BSW、BS_W、BLW 和 BL_W。为了便于对比，我们在表4-8中同时列出了一步模型与两步模型的预测性能，其中一步模型的性能来自表4-3。图4-9与图4-10分别展示了样本 BSW 及 BS_W 的光谱红移与测光红移的散点图以及 Δz 的区间分布情况。从图4-9可以发现，两

表 4-5 不同分类器 5-折交叉验证的平均分类性能。

Table 4-5 The performance of different classifiers for high and low redshift subsamples.

					High redshift			Low redshift			
样本	输入特征	方法	模型参数	Accu.(%)	Prec.(%)	Rec.(%)	F1(%)	Prec.(%)	Rec.(%)	F1(%)	Time(s)
BSW	Pattern I	XGBoost	max_depth=6 n_estimators=200	99.78	92.48	86.10	89.18	99.86	99.92	99.89	124
BSW	Pattern II	CatBoost	depth=6 iterations=1000	**99.79**	92.33	**87.04**	**89.60**	**99.87**	99.93	**99.90**	19
BSW	Pattern III	RF	max_depth=14 n_estimators=500	99.78	**93.91**	84.54	88.98	99.84	**99.94**	99.89	190
BS_W	Pattern IV	XGBoost	max_depth=6 n_estimators=100	99.77	92.30	85.12	88.56	**99.84**	99.92	99.87	46
BS_W	Pattern V	CatBoost	depth=12 iterations=1000	**99.78**	92.87	**85.21**	**88.88**	**99.84**	99.93	**99.88**	130
BS_W	Pattern VI	RF	max_depth=13 n_estimators=100	99.77	**93.88**	82.89	88.04	99.82	**99.94**	99.87	72

表 4-6 不同分类器的验证性能。

Table 4-6 The performance of different classifiers with different training and test samples.

					High redshift			Low redshift			
训练样本	学习方法	输入特征	验证样本	Accu.(%)	Prec.(%)	Rec.(%)	F1(%)	Prec.(%)	Rec.(%)	F1(%)	训练时间 (s)
BSW	CatBoost	Pattern II	BSW	99.96	100	96.27	98.10	99.96	100	99.98	0.25
BS_W	CatBoost	Pattern V	BSW	99.99	100	99.94	99.97	99.99	100	100	0.27
BS_W	CatBoost	Pattern V	BS_W	99.99	100	99.91	99.95	99.99	100	100	0.36
BSW	XGBoost	Pattern I	BSW	100	100	100	100	100	100	100	2
BS_W	XGBoost	Pattern IV	BSW	99.97	99.60	98.00	98.79	99.98	99.99	99.99	1.5
BS_W	XGBoost	Pattern IV	BS_W	99.97	99.68	97.90	98.78	99.98	99.99	99.99	2
BSW	RF	Pattern III	BSW	99.96	100	96.27	98.10	99.96	100	99.98	11
BS_W	RF	Pattern VI	BSW	99.93	99.53	94.16	96.77	99.94	99.99	99.96	3
BS_W	RF	Pattern VI	BS_W	99.92	99.33	93.07	96.10	99.93	99.99	99.96	3

表 4-7 CatBoost、XGBoost、RF 分别在高、低红移样本上的红移估计性能。

Table 4-7 The performance of photometric redshift estimation on different subsamples with different methods.

样本	输入特征	学习算法	模型参数	MSE	MAE	Bias	σ_{NMAD}	$\sigma_{\Delta z}$	R^2	$\delta_{0.3}$(%)	O(%)	时间 (s)
BSW_H	Pattern I	XGBoost	max_depth=6 n_estimators=500	0.0989	0.1515	-0.0254	**0.0253**	0.3144	0.4777	99.11	2.56	2
BSW_H	Pattern II	CatBoost	depth=7 iterations=3000	**0.0756**	**0.1439**	-0.0055	0.0256	**0.2750**	**0.5910**	**99.72**	**2.17**	5
BSW_H	Pattern III	RF	max_depth=11 n_estimators=1000	0.0781	0.1442	0.0035	**0.0253**	0.2796	0.5804	99.55	2.4	14
BSW_L	Pattern I	XGBoost	max_depth=10 n_estimators=700	0.1526	0.2646	-0.0017	0.1012	0.3907	0.6408	93.77	20.64	32
BSW_L	Pattern II	CatBoost	depth=13 iterations=3000	**0.1497**	**0.2624**	-0.0005	**0.1002**	**0.3870**	**0.6475**	**93.85**	**20.37**	285
BSW_L	Pattern III	RF	max_depth=15 n_estimators=500	0.1536	0.2656	0.0010	0.1012	0.3919	0.6386	93.67	20.68	1300
BS_W_H	Pattern IV	XGBoost	max_depth=5 n_estimators=1000	0.0800	0.1429	-0.0068	0.0250	0.2828	0.5119	99.51	**2.24**	3
BS_W_H	Pattern V	CatBoost	depth=6 iterations=2000	**0.0759**	**0.1404**	-0.0081	0.0248	**0.2756**	**0.5407**	**99.64**	2.32	5
BS_W_H	Pattern VI	RF	max_depth=11 n_estimators=300	0.0769	0.1405	-0.0007	**0.0242**	0.2773	0.5285	99.51	2.59	5
BS_W_L	Pattern IV	XGBoost	max_depth=10 n_estimators=1000	0.1791	0.2947	-0.0002	0.1109	0.4233	0.5967	92.72	23.66	370
BS_W_L	Pattern V	CatBoost	depth=13 iterations=3000	**0.1775**	**0.2934**	-0.0001	**0.1105**	**0.4213**	**0.6005**	**92.80**	**23.52**	450
BS_W_L	Pattern VI	RF	max_depth=15 n_estimators=1000	0.1812	0.2963	0.0006	0.1115	0.4257	0.5919	92.60	23.76	2811

步模型的离群数量要明显少于一步模型，尤其对于高红移区域。因此，由表4-8、图4-9与图4-10可以看出，两步模型要优于一步模型。

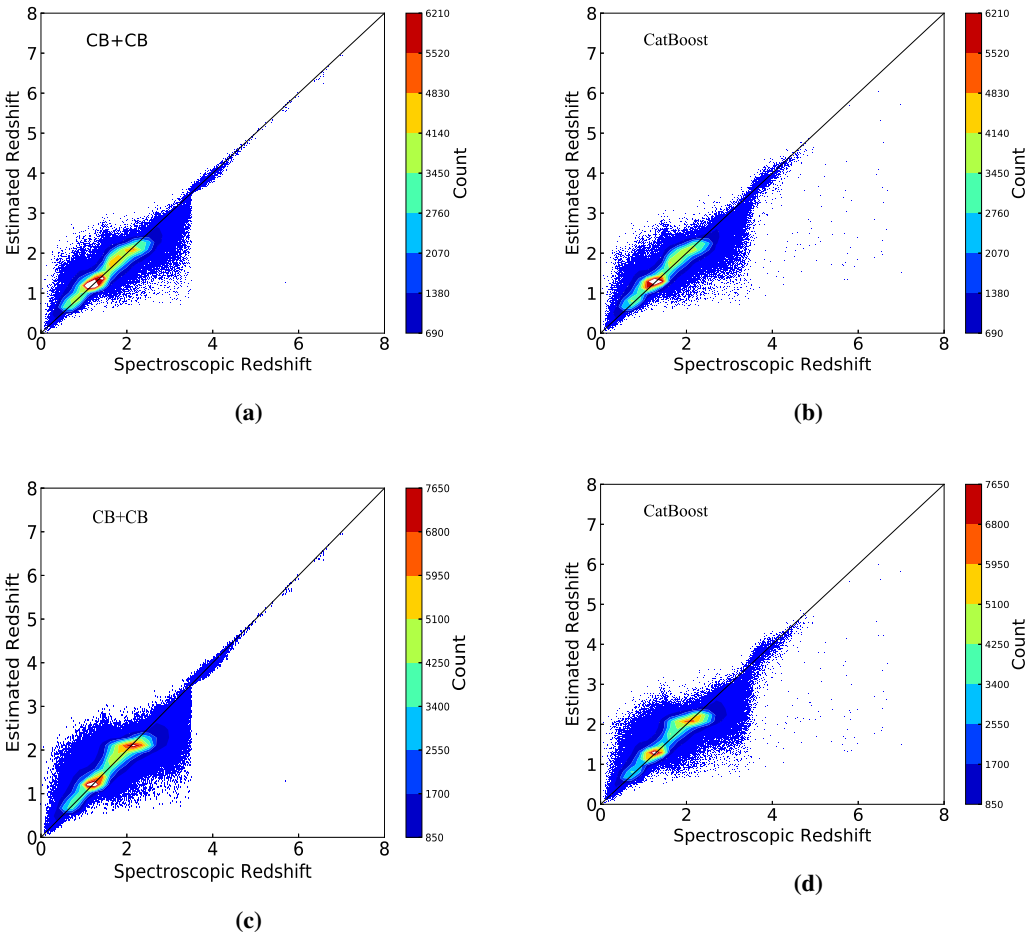


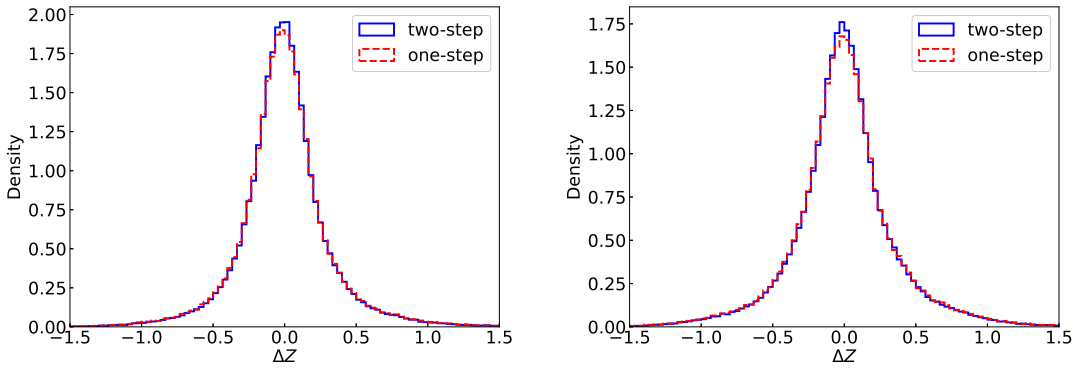
图 4-9 两步模型与一步模型分别在样本 BSW 与 BS_W 上的光谱与测光红移的散点图，其中 (a)、(b) 为样本 BSW 的实验结果，(c)、(d) 为样本 BS_W 的实验结果，(a)、(c) 为两步模型的实验结果，(b)、(d) 为一步模型的实验结果。

Figure 4-9 Comparison of the photometric redshift with the spectroscopic redshift for the samples BSW in panels (a) and (b) and BS_W in the panels (c) and (d) with two-step models (panel a and c) and one-step models (panel b and d), respectively.

表 4-8 两步模型与一步模型的性能对比。

Table 4-8 Comparison of the performance of photometric redshift estimation by two-step model with that by one-step model.

测试样本	MSE	MAE	Bias	σ_{NMAD}	$\sigma_{\Delta z}$	R^2	$\delta_{0.3}$(%)	O(%)	预测时间 (s)
two-step model									
BSW	0.0970	0.2153	-0.00004	0.0854	0.3114	0.7967	96.32	15.13	25.0
BLW	0.1216	0.2103	0.0249	0.0784	0.3487	0.7630	94.86	14.81	2.0
BS_W	0.1266	0.2499	-0.00004	0.0955	0.3558	0.7440	96.32	18.67	35.0
BL_W	0.1251	0.2157	0.0280	0.0804	0.3537	0.7553	94.40	15.63	9.0
one-step model									
BSW	0.1059	0.2223	-0.00002	0.0872	0.3254	0.7780	96.01	15.80	0.80
BLW	0.1239	0.2134	0.0265	0.0797	0.3521	0.7585	94.50	15.11	0.03
BS_W	0.1365	0.2578	-0.000007	0.0981	0.3695	0.7239	94.84	19.42	1.00
BL_W	0.1277	0.2188	0.0296	0.0817	0.3574	0.7502	94.40	15.88	0.05



(a) (b)

图 4-10 两步模型与一步模型分别在样本 BSW 与 BS_W 上的 Δz 分布，其中 (a) 为样本 BSW 的实验结果，(b) 为样本 BS_W 的实验结果。

Figure 4-10 Comparison of the Δz distribution for the samples BSW in panels (a) and BS_W in the panels (b), respectively.

4.6 BASS DR3 类星体侯选体的红移预测

我们采用两种模型分别对 BASS DR3 中的类星体侯选体进行了红移估计，表4-9列出了所有需要的预测模型及其详细参数，包括六个回归器与一个分类器。完整的工作流如图4-11所示。BASS DR3 类星体侯选体来自第三章对 BASS DR3 的分类结果。我们选择了所有可能的侯选体，总数为 26,200,778，不同置信区间的具体数量可参考第三章第 4 节内容。

表 4-9 测光红移预测工作流中涉及的预测模型。

Table 4-9 The models used in photometric redshift estimation workflow.

模型名称	学习方法	输入特征	红移范围	是否含红外数据
Regressor 1^{st}	CatBoost	Pattern II	全部	包含
Regressor 2^{nd}	CatBoost	Pattern V	全部	有的含，有的不含
Regressor 3^{rd}	CatBoost	Pattern II	红移 z≥ 3.5	包含
Regressor 4^{th}	CatBoost	Pattern II	红移 z< 3.5	包含
Regressor 5^{th}	CatBoost	Pattern V	红移 z≥ 3.5	有的含，有的不含
Regressor 6^{th}	CatBoost	Pattern V	红移 z< 3.5	有的含，有的不含
Classifier 1^{st}	CatBoost	Pattern III	高红移与低红移部分	有的含，有的不含

根据图4-11的工作流，首先进行一步模型的预测。BASS DR3 类星体侯选体根据是否有红外数据分成两个样本：一个样本数据中只包含光学特征；另一个样本包含了光学与红外特征。对于含有光学与红外特征的样本，我们采用回归器 I 进行预测。而对于只含光学特征的样本，我们采用回归器 II 进行预测，一步模型预测的红移结果保存在字段 $redshift\_p$ 中。然后我们再将两个样本进行合并，合并的表中包含了一步模型的预测结果。之后，我们再进行两步模型的预测，通过分类器 I 再将类星体侯选体分成高红移与低红移两个样本。对于高红移样本，我们采用回归器 III 进行红移预测。而对于低红移样本，我们采用回归器 IV 进行预测，

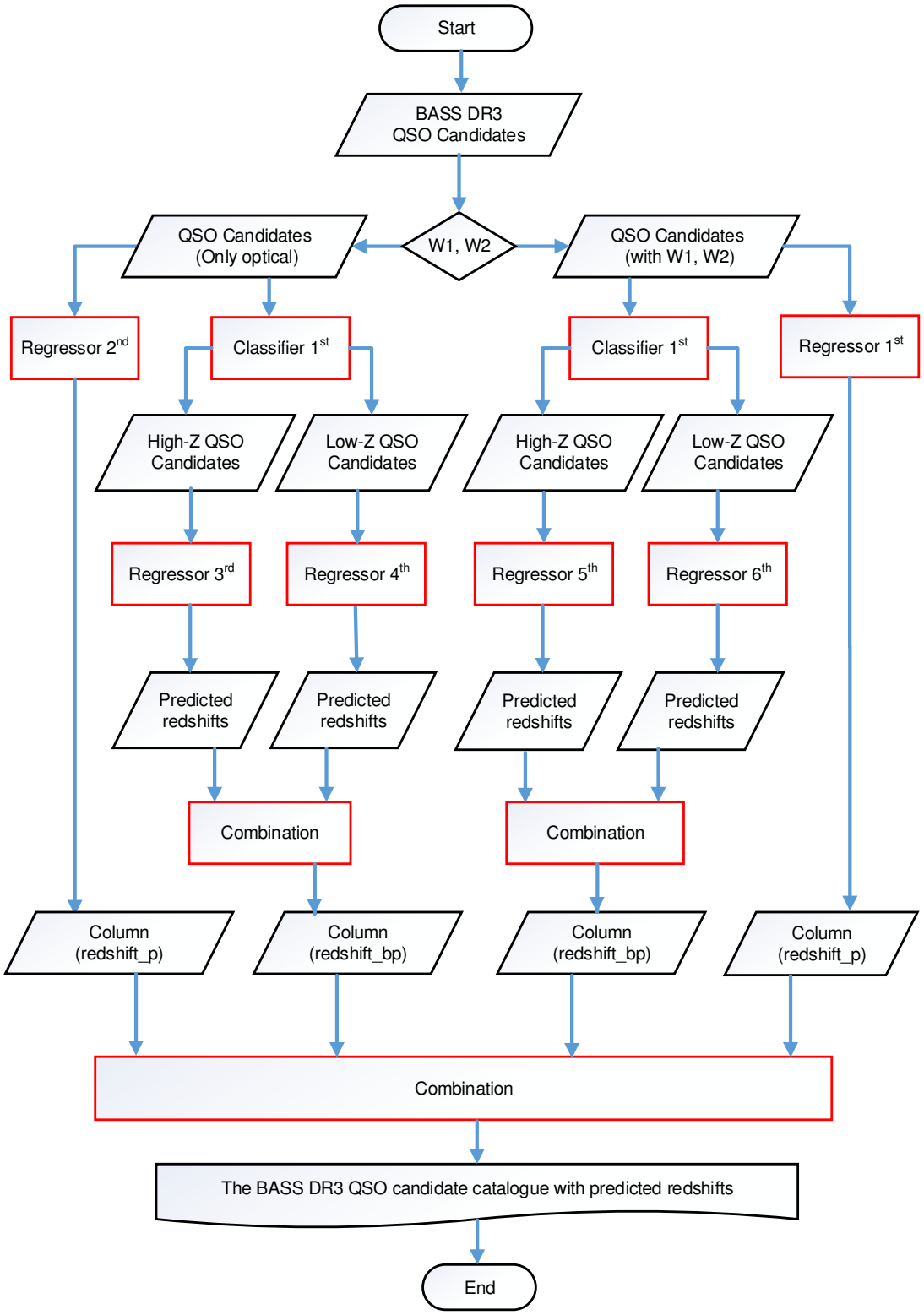


图 4-11 二步模型进行红移预测的工作流程。

Figure 4-11 The photometric redshift estimation workflow by two-step model.

预测结果保存在 *redshift_bp* 字段。最后再将两个样本合并，形成完整的可发布星表，星表中包括了两种方法的测光红移预测数据。考虑到篇幅的限制，表4-10中只列出了部分预测结果，全部的预测结果放在https://doi.org/10.12149/101166。

表 4-10 BASS DR3 类星体候选体的测光红移预测星表示例，其中 redshift_bp **是两步模型预测结果，**redshift_p **是一步模型预测结果。**

Table 4-10 The estimated redshifts of BASS DR3 quasar candidates, redshift_bp **is predicted redshift by two-step model,** redshift_p **is predicted redshift by one-step model.**

ID	RA.	Dec.	redshift_bp	redshift_p
95429001151	133.48449872099638	84.64760058245369	3.612	2.755
95375007162	146.34468863159154	84.19521072349899	3.756	2.967
95375009802	146.70683096395808	84.34011991887981	3.853	3.071
95376013790	156.8204830082884	84.55993969686726	3.820	3.462
95432000953	156.68410784655092	84.62164596197978	3.770	2.216
95433002683	162.41293319964882	84.75661154235415	3.831	3.252
95379009607	172.10642294951734	84.54070384254878	4.290	2.970
95435000350	172.16026329686127	84.59069887892409	3.584	2.644
95439001178	204.45978967988097	84.63147943935314	3.703	2.324
95440003112	208.24299714603515	84.74509303019737	3.835	2.349
95440003256	207.3572024876584	84.75161970723823	3.657	3.329
95441002154	214.02787846450985	84.7184000420548	3.643	2.777
95372005855	128.4637617956736	84.17660965013285	3.690	2.329
95372007632	127.96938333672199	84.25455113312614	3.818	3.120
95372008421	129.9774226081903	84.30134401378302	3.753	2.407
95373007697	136.73203714550777	84.23113544297041	3.875	2.745
95373009831	138.71782967532423	84.32491202628485	3.725	2.288
95374010575	143.50234604548734	84.37099267752663	3.858	3.136
95312008225	151.1039377338386	83.5936910539942	3.716	3.013
95312012685	150.5329880180199	83.78811573097289	3.861	1.869
95376009700	153.0999380695831	84.3246689896127	3.738	2.487

根据两种方法预测的结果及类星体分类时的概率，我们分别统计了不同概率下类星体的红移区间分布。如图4-12可见，大部分候选体的红移分布在 $1 < z < 2$ 之间，小部分候选体的红移大于 3 以上，这种分布曲线与现有的红移分布是相一致的。

表 4-11 概率大于 95% 的类星体候选体在两种红移预测模型下不同测光红移区间的数量。

Table 4-11 The number of BASS DR3 quasar candidates with $P_{\rm Q} > 0.95$ in different redshift ranges by two models.

模型	红移 < 3.5	3.5 ≤ 红移 < 4.5	4.5 ≤ 红移 < 5.5	红移 ≥ 5.5
一步模型	796,078	2,822	27	1
两步模型	794,990	3,817	97	24
一步模型或两步模型	796,100	3,960	125	24
一步模型且两步模型	794,968	2,828	24	1

然后我们对概率大于 95% 的类星体候选体进行了不同预测的测光红移区间的数量统计，各区间的数量详列在表4-11中。从统计数据表明，两步模型在高红

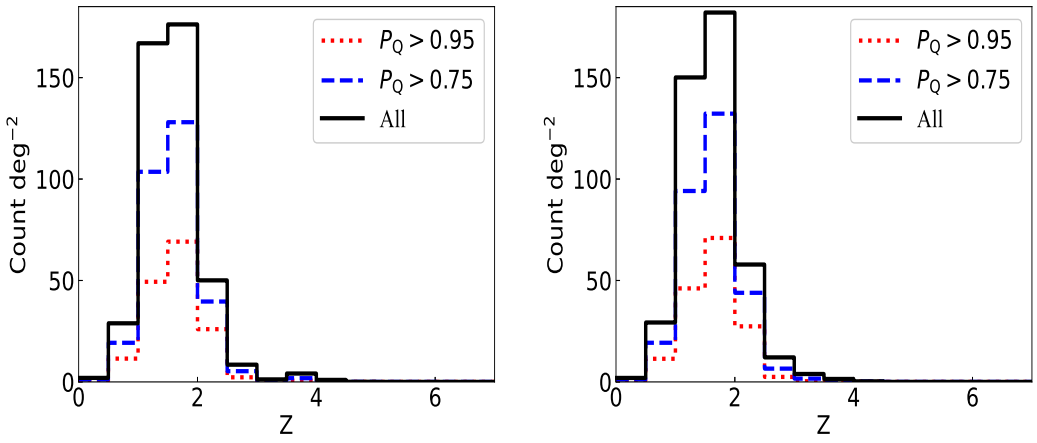


图 4-12 不同概率的 BASS DR3 类星体候选体的估计测光红移区间分布。

Figure 4-12 Left panel: the number density of quasar candidates as a function of photometric redshifts by two-step model; Right panel: the number density of quasar candidates as a function of photometric redshifts by one-step model (Regressor 1^{st}). In both panels, the quasar candidates contain both optical and infrared information with the same predicted results by binary and multiclass classifiers for different probabilities ($P_Q > 0.95$: red dotted line, $P_Q > 0.75$: blue dotted dash line, all: black line).

移区间的数量明显要多于一步模型，这将有助于高红移类星体的发现。表4-12详细列出了所有一步模型或两步模型预测的红移大于 4.5 的类星体候选体。

表 4-12 高红移类星体侯选体。

Table 4-12 The BASS DR3 quasar candidates with $\mathrm{redshift_{bp}}$**> 4.5 or** $\mathrm{redshift_{p}}$**> 4.5.**

RA.	Dec.	g	r	z	W1	W2	redshift_bp	redshift_p
351.205	-0.006	23.126	22.275	22.278	16.882	16.176	6.346	4.086
134.283	32.176	22.908	19.646	18.592	15.681	15.214	4.767	4.578
149.133	32.270	22.604	19.853	19.114	16.051	15.607	4.632	4.559
190.684	32.295	22.435	20.254	19.440	16.606	16.011	4.466	4.535
221.632	32.814	21.324	20.299	19.258	15.834	14.909	5.536	3.563
242.088	32.659	20.135	19.505	18.844	14.716	13.237	5.102	3.172
245.430	32.750	21.885	20.869	19.757	15.971	14.812	5.670	3.379
247.594	32.876	22.429	20.876	20.017	16.227	15.153	5.157	3.912
129.096	33.649	21.246	20.768	20.731	17.187	15.882	5.576	1.695
143.279	33.432	23.454	20.282	19.407	16.314	16.007	4.641	4.530
104.497	34.186	23.591	21.033	19.839	17.307	16.170	4.609	3.257
129.173	33.793	22.336	22.465	21.828	17.126	16.188	6.289	2.832
143.421	34.056	21.697	21.489	20.999	17.053	16.297	6.510	2.049
146.400	33.908	21.688	21.372	21.220	17.275	16.528	5.776	1.993
152.664	34.168	23.305	20.913	19.905	16.265	15.448	4.563	3.079
192.426	33.832	23.397	20.373	18.918	16.227	15.456	4.775	4.107
218.300	34.035	23.305	20.378	20.009	16.488	16.012	4.525	4.240
114.161	34.975	20.349	19.952	19.669	15.509	14.664	5.570	1.716
132.063	34.721	22.644	21.272	19.908	15.613	14.140	5.643	4.294
154.882	34.988	22.974	20.364	19.650	16.556	15.897	4.506	4.490

续表见下页

表 4-12 续表。

RA.	Dec.	g	r	z	W1	W2	$redshift_{bp}$	$redshift_{p}$
232.056	34.715	19.775	19.813	19.892	16.934	15.545	5.126	2.398
252.206	34.743	21.982	21.904	21.576	18.003	17.360	5.787	3.310
254.569	34.469	18.216	18.026	17.832	15.102	13.775	5.107	2.297
196.266	35.330	20.471	20.057	19.517	15.858	14.725	5.071	1.819
247.825	35.342	21.596	21.114	20.685	16.585	15.846	5.911	1.830
255.302	35.432	21.873	20.963	20.149	15.624	14.588	4.527	2.212
125.337	36.136	22.784	20.254	19.329	16.435	15.848	4.598	4.583
133.966	36.087	21.331	20.879	20.628	15.866	15.082	5.625	1.334
167.604	36.003	20.189	19.989	20.050	16.630	15.407	5.726	1.756
218.510	35.990	22.102	21.162	19.988	16.789	15.042	6.941	5.715
242.453	35.850	20.702	20.525	20.271	17.288	16.124	5.765	2.130
126.990	37.035	19.421	18.736	18.466	15.035	13.484	4.722	2.225
225.374	36.531	21.454	20.300	19.387	15.559	14.149	5.338	2.469
180.242	37.211	23.125	20.875	19.754	16.698	16.169	4.685	4.586
182.842	37.901	23.717	20.692	19.587	16.815	16.115	4.687	4.382
194.325	37.792	22.692	19.835	19.033	16.193	15.837	4.703	4.465
218.562	38.361	21.161	20.933	20.953	17.851	16.835	6.459	3.635
224.144	37.789	21.866	21.236	20.353	16.363	14.938	4.818	2.843
257.212	38.260	19.658	19.510	19.958	16.620	15.280	5.146	3.434
264.033	37.838	23.689	20.878	19.726	16.725	16.423	4.711	4.870
137.814	38.620	22.751	21.280	20.015	15.734	14.423	4.869	2.537
274.064	39.686	23.446	20.788	19.944	17.193	16.707	4.630	4.584
144.219	40.263	23.419	20.479	20.090	16.508	16.119	4.550	4.431
244.582	40.041	20.958	20.101	19.465	15.654	14.142	4.924	2.629
256.319	39.701	21.159	21.003	21.048	16.744	15.823	5.892	1.663
151.207	40.765	24.122	21.181	19.763	16.928	16.279	4.770	4.237
160.857	40.814	23.786	20.649	19.106	15.874	15.062	4.769	3.447
220.677	40.602	21.405	20.912	20.279	17.103	15.713	4.507	2.500
104.559	41.107	23.547	22.368	22.387	17.173	16.529	4.533	2.586
212.368	41.327	21.186	20.626	20.361	16.196	14.792	4.661	1.549
255.287	41.326	21.534	20.579	19.812	16.624	15.513	5.571	3.047
175.801	42.195	22.613	20.456	19.675	17.048	16.209	4.552	4.519
256.399	42.535	21.057	20.565	19.946	15.745	14.504	4.595	1.910
150.896	43.126	23.019	20.766	19.689	16.224	15.273	4.754	3.315
193.152	43.016	24.110	20.692	20.026	16.918	16.346	4.581	4.323
269.101	43.181	22.997	20.121	19.377	16.594	16.237	4.671	4.606
225.116	43.700	21.944	19.383	18.738	15.968	15.486	4.619	4.506
185.175	44.705	22.819	20.510	19.603	16.633	16.059	4.545	4.571
202.105	44.750	23.921	21.030	19.936	16.248	15.507	4.608	3.741
229.256	44.341	22.961	20.856	19.119	13.742	12.346	5.277	1.636
236.896	44.781	22.860	20.568	19.546	16.469	15.980	4.589	4.402
112.763	44.997	23.802	20.329	18.756	15.824	15.300	4.934	4.520
256.581	45.465	21.022	20.498	19.918	16.062	14.595	5.274	2.787
253.398	45.930	23.557	20.470	19.462	16.265	15.790	4.651	4.626
261.382	46.271	22.888	20.216	19.275	16.370	16.040	4.617	4.504
203.209	46.852	23.904	20.549	19.431	16.077	15.385	4.557	3.209
253.583	46.321	23.163	20.477	19.664	17.067	16.429	4.537	4.518
131.394	47.045	23.604	22.345	22.453	17.134	16.684	4.547	2.621
194.982	47.283	23.644	20.557	19.797	16.455	15.857	4.539	4.253

续表见下页

表 4-12 续表。

RA.	Dec.	g	r	z	W1	W2	$redshift_{bp}$	$redshift_{p}$
228.520	47.638	23.085	20.754	19.742	16.745	16.382	4.614	4.594
123.578	48.595	22.774	20.446	19.662	16.759	16.252	4.490	4.584
194.317	48.330	22.421	19.789	19.256	16.348	15.954	4.566	4.459
147.250	49.539	23.432	20.921	19.821	16.215	15.796	4.545	4.304
169.759	49.525	20.006	19.108	18.909	15.417	14.154	4.634	2.106
229.330	49.001	23.046	20.265	19.259	16.600	16.231	4.676	4.692
218.205	49.982	20.846	20.419	20.575	16.511	15.314	6.935	1.517
234.209	50.136	23.003	19.928	18.315	15.128	14.520	4.894	4.601
265.605	50.032	23.866	20.467	19.445	16.171	15.755	4.643	4.380
185.129	50.840	23.685	21.109	19.912	16.543	15.910	4.523	3.676
280.928	50.451	22.884	20.120	18.907	15.916	15.438	4.654	4.292
244.068	51.560	22.333	19.545	19.165	15.950	15.609	4.523	4.285
288.622	51.355	22.664	20.455	19.390	16.345	15.932	4.576	4.307
227.348	51.721	23.109	20.511	19.551	16.500	16.158	4.488	4.613
185.568	52.647	22.931	20.382	19.534	16.291	15.801	4.519	4.431
196.744	52.707	21.380	20.884	20.072	16.145	15.235	4.573	2.525
142.080	53.673	23.589	19.763	18.464	15.156	14.627	4.671	4.389
169.309	54.155	23.962	20.895	19.875	16.708	16.501	4.612	4.310
224.310	53.720	20.977	20.705	20.458	16.867	15.748	6.328	1.890
231.208	54.040	20.841	20.554	20.322	16.188	15.172	5.316	1.717
233.961	53.728	21.483	21.146	21.145	17.629	16.390	6.545	1.955
241.894	53.706	20.301	19.407	18.836	15.563	14.338	4.751	2.499
190.627	54.383	23.364	20.423	19.652	16.212	15.688	4.642	4.458
242.266	54.644	21.872	21.355	21.840	16.855	15.587	6.353	3.974
276.157	54.690	23.729	20.329	19.234	16.382	15.766	4.769	4.566
239.431	55.067	23.787	22.510	22.320	17.144	16.446	4.568	2.065
150.759	55.614	21.646	21.056	20.343	16.326	15.009	4.874	2.182
166.619	55.682	23.303	20.212	19.456	16.389	15.727	4.575	4.477
226.878	55.620	22.928	19.890	19.339	15.969	15.570	4.538	4.345
269.518	56.179	23.886	20.714	20.183	17.048	16.552	4.550	4.216
121.416	56.817	23.618	21.093	19.983	16.591	16.276	4.576	4.072
140.904	58.008	21.102	20.117	19.763	15.562	14.551	4.838	2.782
205.681	58.647	23.968	21.256	19.959	17.218	16.617	4.723	4.058
195.574	59.467	21.866	21.747	21.483	17.704	16.859	6.010	2.382
198.837	59.076	21.741	20.881	20.461	16.511	15.413	6.252	2.308
232.089	59.180	23.136	20.695	19.836	16.882	16.420	4.491	4.592
145.285	59.790	24.018	20.548	19.415	16.441	15.877	4.725	4.603
126.623	60.926	22.574	20.909	20.244	16.735	15.642	4.582	3.124
194.712	61.961	23.936	20.237	19.486	16.090	15.629	4.635	4.098
268.876	61.712	23.699	20.414	19.697	16.652	16.489	4.625	4.326
137.219	63.382	23.859	22.581	22.730	17.277	16.566	4.876	2.188
203.227	62.888	23.472	20.161	19.526	16.168	15.657	4.596	4.311
257.693	64.003	24.112	21.155	20.051	16.791	16.306	4.719	4.250
152.723	64.809	23.174	20.477	19.352	16.377	15.928	4.674	4.677
201.775	65.785	23.904	20.793	19.935	16.642	16.038	4.584	4.176
123.384	66.655	23.143	20.348	19.949	16.556	16.236	4.557	4.360
201.749	67.378	23.370	21.275	19.616	15.327	13.838	5.406	2.798
248.683	67.575	22.865	20.345	19.096	16.073	15.493	4.645	3.642
209.675	68.986	23.611	20.517	19.837	16.593	16.167	4.582	4.345

续表见下页

表 4-12 续表。

RA.	Dec.	g	r	z	W1	W2	$redshift_{bp}$	$redshift_{p}$
283.249	69.332	23.047	20.383	19.597	16.763	16.477	4.595	4.618
184.502	70.148	23.103	20.479	19.848	17.040	16.825	4.523	4.294
148.458	73.753	22.319	19.993	18.965	16.161	15.215	4.542	3.331
139.314	74.732	22.915	19.642	19.022	15.658	15.235	4.548	4.439
217.540	76.814	23.225	20.266	19.583	15.984	15.444	4.502	4.304
137.349	82.513	23.312	20.329	19.577	16.584	16.354	4.602	4.506
161.075	84.577	23.486	20.843	19.969	17.073	16.408	4.524	4.420

4.7 本章小结

本章主要介绍了利用机器学习算法进行类星体的测光红移估计的研究。对比了 CatBoost、XGBoost 和随机森林算法在测光红移估测上的性能。通过实验证明，CatBoost 算法与 XGBoost 算法总体性能相当，但对于本样本而言，CatBoost 稍有优势，但在计算速度上，CatBoost 要优于 XGBoost 与随机森林。我们设计了两种测光红移估计的方案：一步模型与两步模型。一步模型比较简单而直接，针对全样本进行训练，构建一个全局的预测模型；两步模型则采用先分类，即先将类星体分为高红移与低红移类星体两大类，然后再针对不同部分分别构建回归模型。最后采用这两种方法对前一章预测的 BASS DR3 类星体侯选体进行了红移估计，其中任意一种方法预测的红移大于 3.5 的共有 3,960 个，而红移大于 4.5 的类星体侯选体有 125 个，为后续进行的光谱巡天提供了类星体侯选源。

第 5 章　星系的测光红移

暗能量巡天之图像巡天计划包括了三个子巡天，分别是暗能量相机多色巡天（the Dark Energy Camera Legacy Survey, DECaLS)，北京-亚历桑那巡天（the Beijing-Arizona Sky Survey, BASS）和梅奥 z 波段巡天（the Mayall z-band Legacy Survey，MzLS)，在不到 3 年的时间里合并观测了近 14,000 平方度的天区，包括 g、r、z 三个光学波段的图像。2019 年，DESI 图像巡天发布了最新一版的数据（DR9)，包括测光图像及星表，其中星表不仅包括了 g、r、z 三个光学波段，还包括了由 unWISE 红外图像获取的 4 个红外波段的测光信息。所有测光流量数据由测光软件 'TRACTOR' 从原始图像提取得到。整个星表包括了近 20 亿个天体，其中星系有 11 亿。通过对如此大规模的星系进行红移预测，可以建立宇宙的三维空间图像，从而对于宇宙学的膨胀历史和暗能量的本质研究都具有重要意义。

5.1　样本数据

对于已知星系样本的红移，我们采用来自 SDSS DR16 及 LAMOST DR7 中的星系光谱红移。DESI 网站发布了 SDSS DR16 与 DESI DR9 的星表按 1.5 角秒交叉后的结果星表，包括了光谱的 MJD、PLATE 和 FIBERID 字段。我们将这个结果星表与 SDSS DR16 的星系表进行联合查询，就得到了每个星系的光谱红移，这个已知样本称为 DSW，我们同样按 1.5 角秒进行 LAMOST DR7 与 DESI DR9 的交叉，每个源只保留最近的一个对应目标，得到的已知结果星表称为 DLW。然后我们再对样本 DSW 和 DLW 进行如下操作：

(1) 将样本 DSW 和 DLW 中的流量（包括模型流量及孔径流量）数据转换为星等，计算公式为 $M = 22.5 - 2.5lg(F)$，F 为流量，M 为对应的星等。

(2) 去除无效源。无效源是指那些超出极限星等范围（$g > 24, r > 23.4$ 或 $z > 22.5$) 和测光数据差（maskbits!=0) 的数据。此外，对于样本 DSW，我们要求 z_warning=0, 而对于样本 DLW，要求 $z\_err < 0.01$。

图5-1为样本 DSW 和 DLW 的红移分布直方图，样本 DSW 的光谱红移范围为 0 到 2，而样本 DLW 的光谱红移范围为 0 到 1。由图5-1所示，已知样本在 r 波段星等大于 22 等时及高红移样本数量明显偏少。因此，我们采用模板匹配与机器学习相结合的方法来进行星系的测光红移估计。

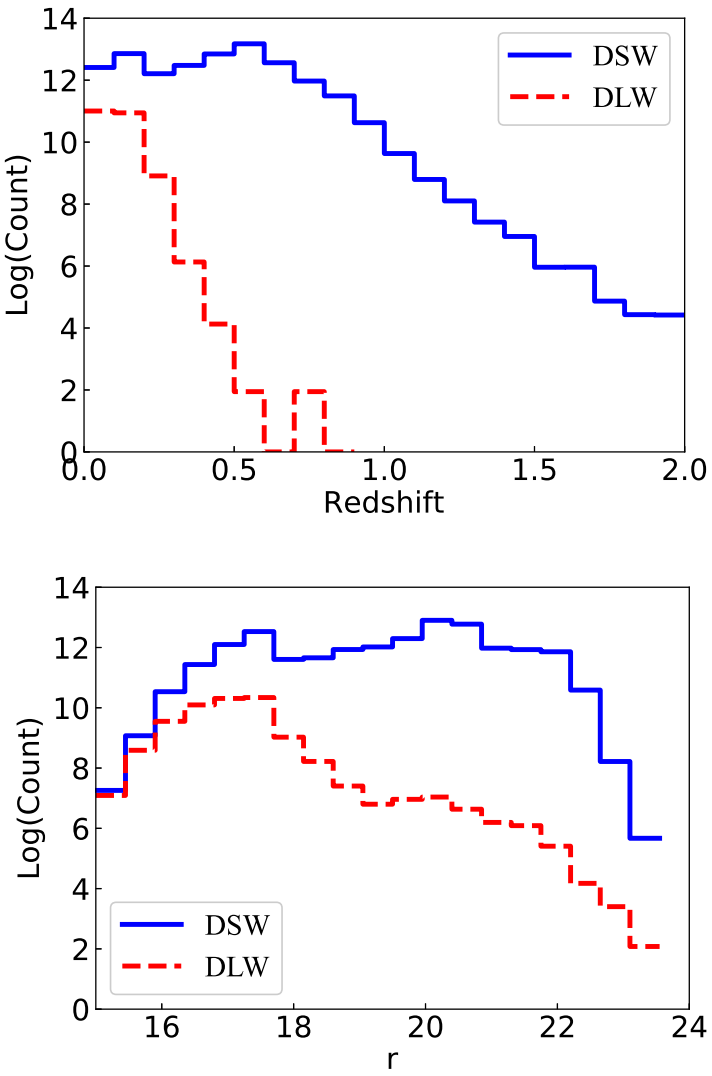


图 5-1 样本 DSW 和 DLW 的红移与 *r* 星等分布直方图。

Figure 5-1 The histogram of redshift and *r*-magnitude for the samples DSW and DLW.

5.2　基于模板匹配方法的预测

基于模板匹配方法估计测光红移的工具软件有很多，比如 LePHARE、BPZ 、HyperZ、EAZY 等。Euclid Collaboration (Euclid Collaboration, 2020) 为了给 Euclid 望远镜选择合适的红移预测方法，对多个方法进行了比较，其中模板匹配方法包括了 LePHARE、CPz、Phosphoros 和 EAZY。EAZY 在多个测试样本上所获得的离群率 O 指标都是最优的，而 σ_{NMAD} 指标值与最好的 LePHARE 方法给出的值也非常接近。Schmidt 等 (Schmidt et al., 2020) 为满足 LSST 的需求也做了类似的工作。他们选择的模板匹配方法包括 BPZ、LePHARE 和 EAZY。在相应样本上获得的 PIT (The Probability Integral Transform) 和 CDE (The Conditional Density Estimation) 性能指标如表5-1所示，EAZY 方法表现出优于其他两种方法的性能。综合考虑各项指标，我们选择 EAZY 作为主要的模板匹配方法来进行星系的红移估计。

表 5-1 LSST 模拟数据中采用的三种模板匹配方法进行红移估计时的性能对比 (Schmidt et al., 2020)。

Table 5-1 Performance comparison of three template fitting methods for redshift estimation on LSST simulation data.

方法	PIT 指标	CDE 指标
LePHARE	0.0486	-1.66
BPZ	0.0192	-7.82
EAZY	0.0154	-7.07

EAZY(Brammer et al., 2008) 是一个采用模板匹配方法进行红移估计的开源代码。基于光谱能量分布模板，建立红移与流量、颜色等的关系网格。然后再基于公式 5-1，找到匹配最好的红移，使得 χ^2 的值最小。

$$\chi^2_{z,i} = \sum_{j=1}^{N_{filt}} \frac{(T_{z,i,j} - F_j)^2}{(\delta F_j)^2} \tag{5-1}$$

N_{filt} 波段数量, $T_{z,i,j}$ 是指模板 i 中对应红移 z 与波段 j 的合成流量, F_j 是波段 j 的实测流量，δF_j 是对应实测流量 F_j 的误差。

此外，EAZY 算法可以将多个模板联合起来进行预测。公式 5-2 是多个模板使用时的联合计算方法，其中 α 是指最好的相关系数。

$$T_z = \sum_{i=1}^{N_{temp}} \alpha_i T_{z,i} \tag{5-2}$$

这种多模板的联合对于准确性具有巨大的改进，但也将花费更多的计算时间。此外，EAZY 在进行每一次红移估计时，会计算一个红移可靠性参数 Q_z，计算方法如公式 5-3 所示。一般来讲，对于每个红移估计值，如果 $Q_z < 1$，则认为结果

是可靠的。

$$Q_{\mathrm{z}}=\frac{\chi^2}{N_{\mathrm{filter}}-3}\frac{z_{\mathrm{up}}^{99}-z_{\mathrm{lo}}^{99}}{p_{\Delta z=0.2}} \tag{5-3}$$

式中, χ^2, N_{filter} 与公式 5-1 意义相同, $p_{\Delta z}$ 表示处于测光红移估计值的 $\pm\Delta z$ 范围内的估测数据占全部估测数据的比例。

我们首先采用 EAZY 方法对训练样本 DSW 进行红移预测。在本实验中，我们使用最新版本的模板库，模板库中包含了 9 个模板文件。我们设置参数 $TEMPLATE\_COMBOS=a$,表示联合使用所有模板,参数 N_MIN_COLORS =4, 表示最少要求有 4 个波段的信息，其它参数采用缺省值。然后，我们再根据望远镜在各波段的响应曲线，修改 EAZY 中对应的曲线文件。因为 DECaLS 巡天与 BASS&MzLS 巡天在 g,r,z 三个光学波段的响应曲线都不相同，所以我们把样本 DSW 分成南北两部分数据分别预测。表5-2显示了两个子样本的预测性能。表5-3为去除了 $Q_{\mathrm{z}}>1$ 的预测结果后再计算的性能。对比表5-2和表5-3，去掉 $Q_{\mathrm{z}}>1$ 的部分后，性能有了很大的改进。Q_{z} 反映预测结果的好坏，其数值越大，表示预测效果越差。图5-2展示了基于 EAZY 方法的光谱红移与测光红移的散点图和 Δz(norm) 的分布图。

表 5-2 EAZY 测光红移预测的性能。

Table 5-2 The performance of photometric redshift estimation with EAZY.

子样本	MSE	MAE	*Bias*	σ_{NMAD}	$\sigma_{\Delta z(\mathrm{norm})}$	$\delta_{0.3}(\%)$	$O(\%)$
DSW_south	0.0301	0.0719	-0.0217	0.0243	0.1128	98.79	5.12
DSW_north	0.0341	0.0621	0.0104	0.0224	0.1283	99.18	3.43

表 5-3 满足 $Q_{\mathrm{z}}<1$ 条件时，EAZY 测光红移预测的性能。

Table 5-3 The performance of photometric redshift estimation by EAZY when $Q_{\mathrm{z}}<1$.

子样本	MSE	MAE	*Bias*	σ_{NMAD}	$\sigma_{\Delta z(\mathrm{norm})}$	$\delta_{0.3}(\%)$	$O(\%)$	$Q_{\mathrm{z}}>1(\%)$
DSW_south	0.0088	0.0521	-0.0149	0.0219	0.0539	99.73	2.06	10.43
DSW_north	0.0085	0.0464	0.0044	0.0209	0.0574	99.83	1.77	6.56

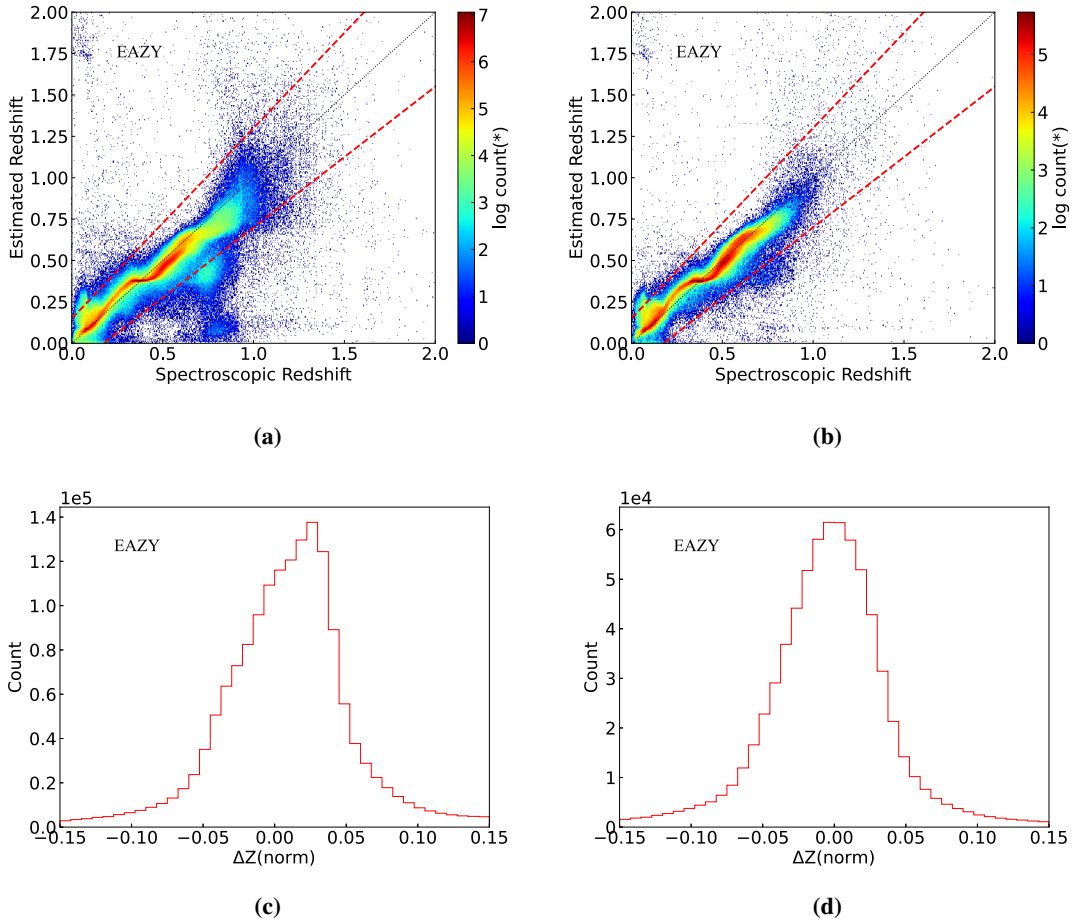


图 5-2 基于 EAZY 方法的测光红移与光谱红移的散点图（图 a 和 b）和 Δz(norm) 的分布图（图 c 和 d）。

Figure 5-2 The scatter figure and Δz(norm) distribution of estimated photometric redshifts and spectroscopic redshifts for the subsamples of the DSW sample. In the scatter figure, the red dashed line represents Δz(norm) = ±0.3, separately.

5.3 基于 CatBoost 的红移预测

与类星体相比，星系具有更大的样本数量，尤其在较低红移及星等偏亮区域。因此，我们同时也采用机器学习方法对星系的红移进行预测。

5.3.1 预测模型构建

机器学习方法的第一步是选择最优输入特征。样本特征除基本的光学、红外特征外，本次实验我们同时采用了 5 个波段对应的孔径特征。DESI 图像巡天数据 DR9 的星表数据中，每个光学波段包含了 8 个孔径流量，孔径大小分别为 0.5、0.75、1.0、1.5、2.0、3.5、5.0 及 7.0 角秒，红外波段包含了 5 个孔径流量，孔径大小分别为 3、5、7、9 及 11 角秒，这些流量我们同样也转换成了 AB 星等。为了对比孔径特征对红移估计的影响，我们分两种情况进行实验。我们首先在输入特征中不包括孔径特征，通过特征重要性评价及最优输入特征的组合实验，得到最优输入特征，我们定义为 Pattern I。然后，我们使用孔径特征作为输入特征，得到的最优输入特征定义为 Pattern II。最后，将所有特征合并一起，采用同样的方法进行最优输入特征实验，得到的最优输入特征，我们定义为 Pattern III。表5-4列出了 CatBoost 采用默认模型参数时 Pattern I 、II 和 III 作为输入特征分别获得的最优模型性能。

表 5-4 CatBoost 采用默认模型参数时 Pattern I、II 和 III 作为输入特征分别获得的最优模型性能。

Table 5-4 The performance of photometric redshift estimation by CatBoost with default model parameters.

样本	输入特征	MSE	MAE	Bias	σ_{NMAD}	$\sigma_{\Delta z(\mathrm{norm})}$	$\delta_{0.3}(\%)$	$O(\%)$	时间 (s)
DSW	Pattern I	0.0038	0.0304	1.0×10^{-6}	0.0180	0.0406	99.72	1.08	88
DSW	Pattern II	0.0056	0.0412	2.85×10^{-4}	0.0250	0.0504	99.64	2	154
DSW	Pattern III	0.0034	0.0291	3.6×10^{4}	0.0173	0.0384	99.76	0.95	143

从表5-4可以发现，输入特征为 Pattern III 时，性能远高于 Pattern II，这表明模型（model）相关的特征与孔径相关的特征合并时，得到的性能是最优的，孔径特征对于星系红移估计有一定的帮助。因此，我们继续在 Pattern I 和 Pattern III 上进行超参数的优化。我们同样采用网格搜索的方法在 $depth$=[3-15]，$iterations$=[1000-5000] 范围内选择最优参数。每次训练采用 5 折交叉认证的方法获得平均的性能数据。图5-3表示不同 $depth$ 时，性能指标 MSE、O 及 σ_{NMAD} 的变化。表5-5列出了详细的性能数据。

比较 Pattern I 与 III 的最优性能，采用 Pattern III 作为输入特征时获得的性能是最优的，从而进一步表明孔径特征对提高红移预测的性能是有帮助的。我们最终选择 Pattern III 作为构建回归模型的输入特征。

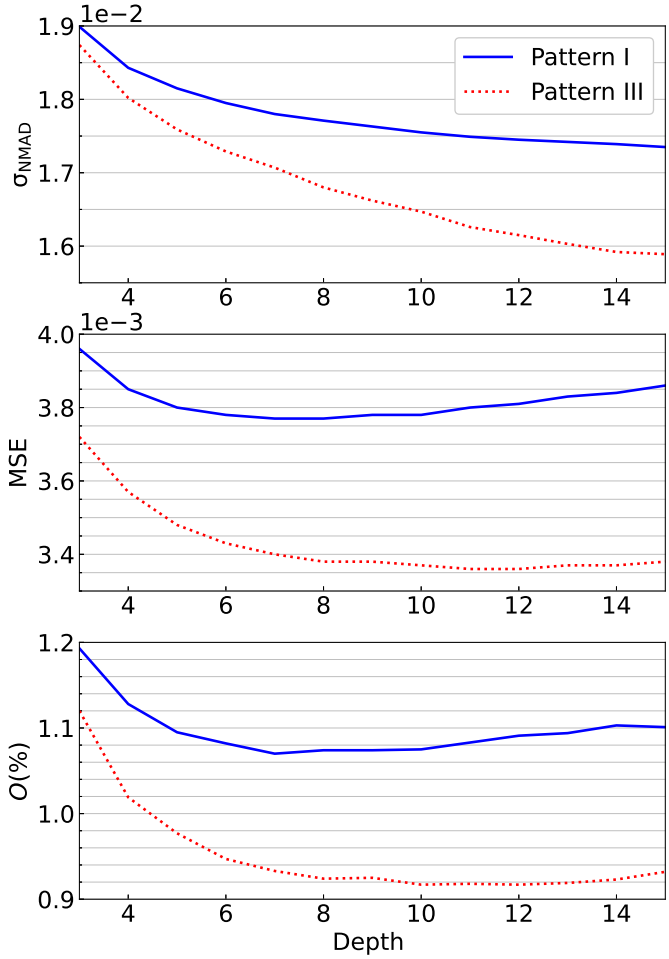


图 5-3 性能指标 MSE、σ_{NMAD} **和** O **随超参数** depth **变化的曲线图。**

Figure 5-3 MSE, σ_{NMAD} **and** O **with different** depth**.**

表 5-5 CatBoost 在训练样本上的最优性能。

Table 5-5 The performance of photometric redshift estimation with the best features and optimal model parameters by CatBoost.

样本	输入特征	模型参数	MSE	MAE	Bias	σ_{NMAD}	$\sigma_{\Delta z(\mathrm{norm})}$	$\delta_{0.3}(\%)$	$O(\%)$	时间 (s)
DSW	Pattern I	$depth = 8$ $iterations = 4000$	0.0037	0.0299	4.8×10^{-5}	0.0175	0.0405	99.72	1.06	381
DSW	Pattern III	$depth = 12$ $iterations = 5000$	0.0032	0.0272	3.3×10^{-4}	0.0156	0.0371	99.76	0.88	3,484

5.3.2 模型验证与讨论

基于最优输入特征及模型参数，再以样本 DSW 进行训练，从而建立回归预测器。利用构建的预测器对训练样本 DSW 进行红移预测，最优性能为 MSE=0.0009, MAE=0.0181, σ_{NMAD}=0.0127, $\sigma_{\Delta z(\mathrm{norm})}$=0.0199, $\delta_{0.3}$=99.98% 及 $O = 0.13\%$。图5-4展示了采用样本 DSW 进行自验证时的光谱红移与预测的红移的散点分布图及 Δz(norm) 的分布图。然后，我们再用样本 DLW 作为外部样本进行验证。我们同时也利用 EAZY 方法对 DLW 进行了预测，为此，我们同样将样本分为 DLW_north 和 DLW_south 两部分。所有预测结果的性能如表5-6所示。

表 5-6 验证样本 DLW 分别采用 EAZY 及 CatBoost 回归模型预测的性能，采用 EAZY 方法时要求 $Q_z < 1$。

Table 5-6 The validation performance of EAZY with $Q_z < 1$ and CatBoost for the test sample.

测试样本	方法	MSE	MAE	Bias	σ_{NMAD}	$\sigma_{\Delta z(\mathrm{norm})}$	$\delta_{0.3}$(%)	O(%)
DLW_north	EAZY	0.0354	0.0632	0.0545	0.0321	0.1538	99.25	5.85
DLW_south	EAZY	0.0247	0.0503	0.0399	0.0227	0.1314	99.46	4.95
DLW_north	CatBoost	0.0014	0.0160	0.0040	0.0076	0.0294	99.79	0.79
DLW_south	CatBoost	0.0014	0.0163	0.0035	0.0078	0.0292	99.79	0.81

由表5-6可以发现，我们构建的回归预测模型在样本 DLW 的预测结果与训练样本的性能相似，表明构建的回归模型在同类数据上具备泛化能力。而将训练模型与 EAZY 方法比较，CatBoost 构建的预测模型在所有性能指标上都要优于 EAZY。但是即便如此，我们并不能说机器学习方法可以完全代替模板匹配方法。机器学习的结果依赖于训练样本，在已知样本足够多和已知红移范围内进行性能评估，机器学习方法具有一定的优势。然而，随着巡天观测越来越深，我们观测的天体也越来越暗，这些未知的暗弱天体超出了已知样本的红移和星等分布范围，导致预测结果的可靠性具有不确定性。在训练样本中，r 星等大于 23 的只有 602 个，z 星等大于 22 等的只有 4,657，占训练样本的比例不到 0.1%，而在 DESI 图像巡天数据 DR9 星表的星系数据中，r 星等大于 23 的数量超过了一半，z 星等大于 22 等的数量也占总数据量的二分之一。由此可见，训练样本无法完全代表实际观测样本，尤其对于暗弱天体，这种情况下用模板匹配方法就比较合适了。因此，我们采用两种方法分别对 DESI 图像巡天数据 DR9 中的星系进行红移预测。

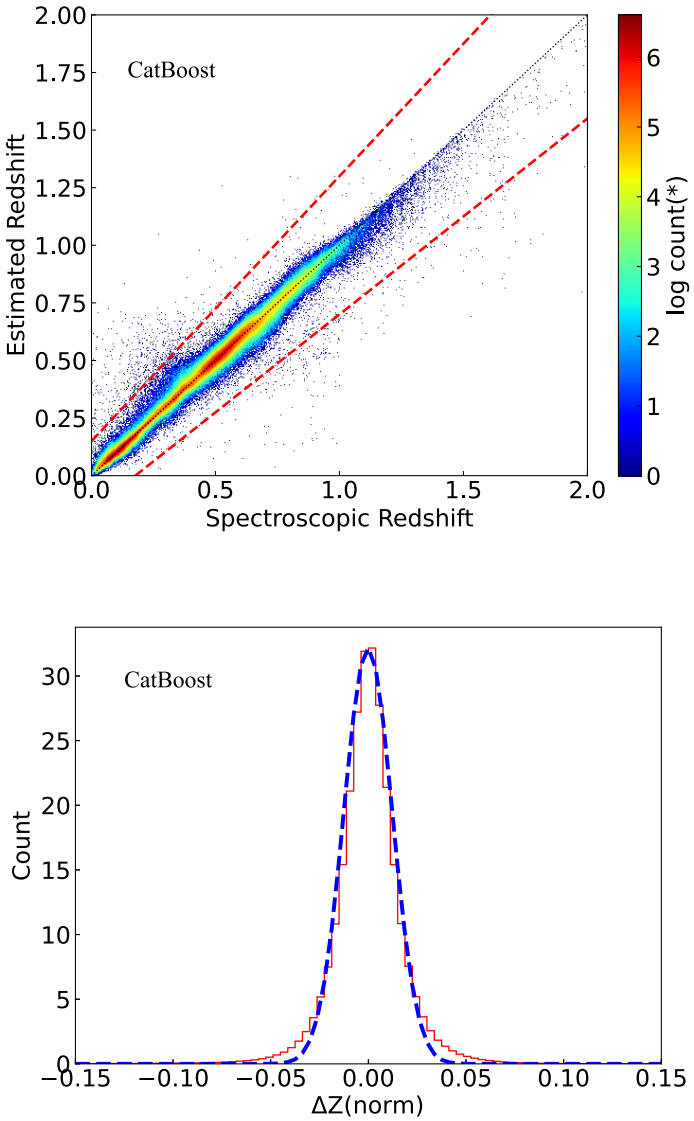


图 5-4 回归模型采用 DSW 验证时的光谱红移与预测红移散点分布图及 Δz(norm) 的分布图。散点图中的红色点线分别表示 Δz(norm) = ±0.3, Δz(norm) 图中的蓝色点线表示高斯分布曲线。

Figure 5-4 The evaluation performance of photometric redshift estimation with CatBoost. In the scatter figure of photometric redshifts vs. spectroscopic redshifts, the red line represents Δz(norm) = ±0.3 separately. For the Δz(norm) distribution, blue dotted curves show the corresponding Gaussian distributions.

5.4 模型应用

我们分别采用 EAZY 和 CatBoost 回归模型对 DESI DR9 中的星系进行了红移估计。DESI DR9 中星系选择的条件为形态分类为‘REX’、‘EXP’、‘DEV’或者‘SER’，总数量为 1,160,568,989。我们对预测程序进行了并行优化（并行化方法参考第六章内容），采用国家天文科学数据中心的高性能计算集群完成了所有的预测计算。受限于已知样本，采用机器学习方法预测的红移，结果都小于 2，而 EAZY 方法预测的红移值在 0 到 6 之间。然后我们对符合 $maskbits = 0$ 及 $Q_z < 1$ 的数据进行了统计，红移区间统计的结果如表5-7所示。

表 5-7 DESI DR9 中星系在不同测光红移区间的数量。

Table 5-7 The number of predicted DESI DR9 galaxies with $maskbits = 0$ and $Q_z < 1$ in different redshift ranges by EAZY.

方法	0<redshift < 2	2 ≤ redshift< 3.5	3.5 ≤redshift< 4.5	4.5 ≤redshift< 5.5	redshift≥ 5.5
EAZY	348,853,174	15,826,023	3,328,406	1,635,656	127,971

我们合并了两种方法的预测结果，形成了一个包含了 DESI 图像巡天数据 DR9 中所有星系的红移星表。星表部分数据如表5-8所示, 完整星表的下载链接为https://doi.org/10.12149/101162。对于星等在极限星等范围内的数据，我们认为预测结果是准确而可靠的，对于超出范围的数据，也可以作为参考。大规模的星系红移星表可以作为星系、星团和宇宙演化等进一步研究的基础。

表 5-8 DESI DR9 星系测光红移星表示例。*z_cb* 为 CatBoost 模型预测的红移，*z_eazy* 为 EAZY 方法预测的红移，Q_z 表示 EAZY 方法的红移估计质量，nfilt 表示有效波段数

Table 5-8 The estimated photometric redshifts of DESI DR9 galaxies, *z_cb* is predicted redshift by CatBoost, *z_eazy* is predicted redshift by EAZY, Q_z demonstrates photometric redshift estimation quality, nfilt is the number of used filters.

release	brickid	objid	RA	Dec	z_cb	z_eazy	Q_z	$nfilt$
9010	465328	3933	140.605	23.897	0.447	0.277	5.420	5
9010	465328	3935	140.605	23.913	0.980	1.395	8.351	5
9010	465328	3936	140.605	24.087	0.520	0.375	0.047	4
9010	465328	3937	140.605	24.075	0.949	1.206	3.747	5
9010	465328	3938	140.605	24.033	0.871	1.142	4.221	4
9010	465328	3940	140.605	23.998	0.650	1.125	56.022	4
9010	465328	3942	140.606	23.946	0.969	1.677	13.574	5
9010	465328	3944	140.606	23.912	0.889	0.932	2.866	5
9010	465328	3948	140.606	24.028	1.140	1.305	1.543	5
9010	465328	3949	140.606	23.874	0.913	1.804	0.011	5
9010	465328	3954	140.606	23.910	0.844	0.500	0.333	5
9010	465328	3955	140.606	24.007	0.858	0.850	1.715	5
9010	465328	3956	140.606	24.045	0.738	0.711	1.145	4
9010	465328	3957	140.606	24.110	0.824	0.884	1.728	5
9010	465328	3958	140.606	24.106	0.468	2.298	3.353	4
9010	465328	3960	140.606	24.117	0.503	1.766	39.261	4
9010	465328	3964	140.607	24.048	0.605	0.320	0.889	4
9010	465328	3965	140.607	24.057	0.953	1.741	0.882	4
9010	465328	3967	140.607	24.031	0.246	0.297	5.305	4
9010	465328	3968	140.607	24.066	0.483	2.658	10.795	5

5.5　本章小结

本章详细介绍了两种星系测光红移的方法，即模板匹配与机器学习方法在大型巡天数据上的应用。我们仍然选择了 CatBoost 作为主要算法，基于已知训练样本构建了红移预测模型，在相类似的数据上进行检测，性能是相一致的。但是，机器学习算法受限于已知样本的特性，容易将高红移星系估计为低红移星系。新型大型巡天所观测的深度都远超已知样本，对于已知样本不足的暗弱天体，模板匹配方法仍然是不错的选择。

第 6 章　面向类星体选源和测光红移估计的在线科研平台

在现今的大型巡天时代，天文大数据带给天文学家的不仅仅是机遇，更多的是挑战。天文大数据的复杂性直接影响着数据的获取、存储、迁移、备份、融合、处理、计算、分析和挖掘等工作。在小规模数据时代，不是问题的问题，在大数据时代可能是严峻的挑战。为解决这些问题，天文学家与计算机、信息学、统计学等专家在不懈地努力，开发出一系列工具、软件，有针对具体课题的，也有普适的。同时，随着天文数据的持续增长，工具与软件的架构设计也在逐渐发生变化，以适应不断发展的硬件平台与性能要求。

从第三章到第五章，我们详细探论了类星体选源和测光红移的算法与性能优化。在类星体选源或测光红移估计等机器学习应用中，还有两个重要环节：数据准备与模型应用。在这两个环节中，通常使用的数据都是整个巡天数据，而大型巡天数据的规模都是非常巨大的，在第二章已经作了详细的描述。大规模的数据不仅在算法设计上遇到了困难，也在数据存储、传输、计算等多个方面对传统的研究模式提出了挑战 (Cui et al., 2015)。本章主要介绍在基于大型巡天数据进行类星体选源和测光红移估测时，我们在数据处理环境、数据准备、数据融合、模型应用等方面采用的一些并行化设计，以期开发出通用的天文数据在线科研平台，为更多的天文科研工作服务。

6.1　虚拟天文台及相关软件工具

通常，天文观测数据分布在不同的数据中心。为了解决数据分散而导致的数据访问问题，天文界提出了虚拟天文台技术。虚拟天文台就是通过先进的信息技术将全球范围内的研究资源无缝透明连接在一起形成的数据密集型网络化天文学研究的平台。2002 年 6 月，国际虚拟天文台联盟（International Virtual Observatory Alliance, IVOA）正式成立，期望联合各国虚拟天文台，制订天文数据互操作的标准与协议，以实现全球天文数据的互连互通。虚拟天文台的基本特征主要表现为三个方面：第一，虚拟天文台的核心作用是实现天文资源的整合，需要把分散的数据统一起来，形成一个物理上分散，逻辑上统一的平台。通过登录这个平台，天文学家可以方便地寻找到自己需要的数据。其次，虚拟天文台提供基础的工具，从而支持一些通用的数据管理操作。例如不同数据资源的交叉联合，数据文件的读写及可视化等。第三，虚拟天文台整合的不仅仅是天文数据资源，也包括天文服务，如天文计算资源、数据挖掘工具、数据可视化工具、数据存储和发布平台、天文文献等各种资源的整合都属于服务的范畴。而虚拟天文台正是通过整合这些异构多样的服务，形成一个统一的科研环境，并且能为了一项共同的研究而相互协同合作。图6-1展示了国际虚拟天文台联盟设计的虚拟天文台的技术架构。这个架构下包括了数据资源、计算资源等不同的模块，制订了数

据访问协议，并且通过注册实现多种服务资源的扩展与接入。

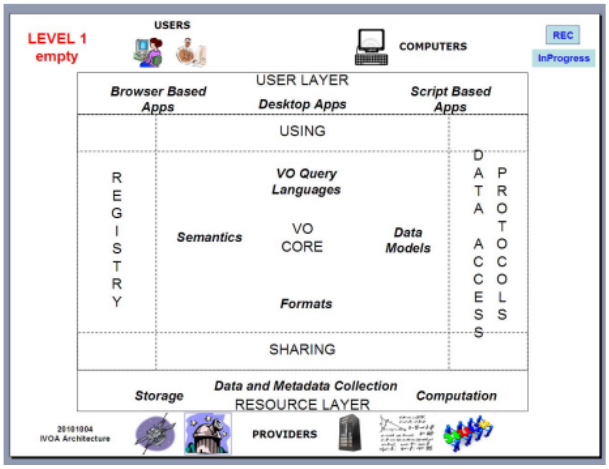


图 6-1 国际虚拟天文台联盟技术架构 (Arviset et al., 2010)。

Figure 6-1 IVOA Architecture Level 1.

法国的斯特拉斯堡数据中心（Strasbourg astronomical Data Center, CDS）[1] 就是利用虚拟天文台技术的一个典型代表。该数据中心汇集了来自各大望远镜的星表数据并通过统一的平台 VizieR 访问 [2]。最新的星表数目达到了 22,747 个，像 Pan-STARRS1、SDSS、Gaia 等主要项目的星表都可以通过此数据中心的平台直接访问和下载。图6-2表示了 VizieR 所有星表覆盖的天区及密度。VizieR 提供了位置、名称等多种星表检索方法，天文学家很容易找到自己想要的数据。

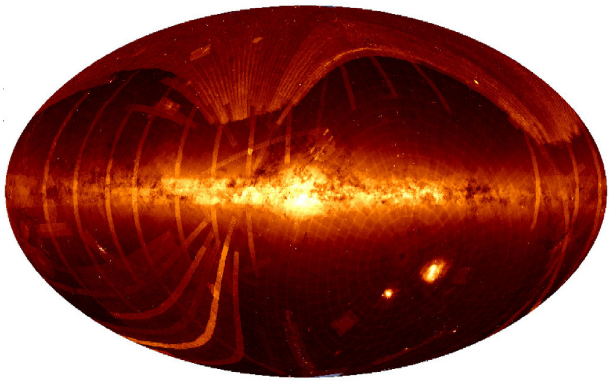


图 6-2 VizieR 的天文数据库覆盖天区情况。

Figure 6-2 VizieR footprint.

CDS 数据平台同时也整合了多种数据访问与处理的工具，像数据可视化工具 ALADIN、数据处理桌面级工具 TOPCAT、数据交叉服务 CROSSMATCH 等。TOPCAT 是一个非常受天文学家欢迎的星表数据处理工具，支持图形界面及命令行两种工作方式。图6-3展示了 TOPCAT 的工作界面。通过 TOPCAT 可以浏览和处理星表数据，并进行多种统计分析、星表交叉融合及可视化等操作。

[1] https://cds.u-strasbg.fr

[2] http://vizier.china-vo.org

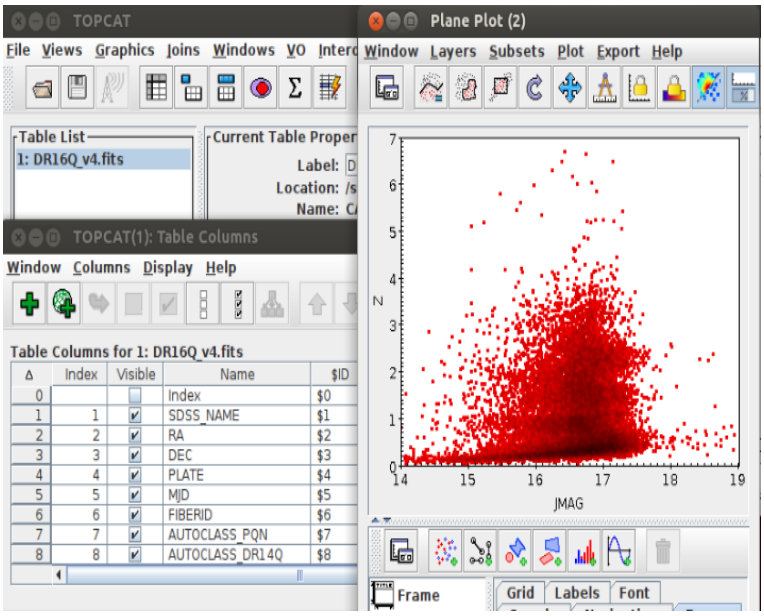


图 6-3 TOPCAT 软件工作界面。

Figure 6-3 The user interface of TOPCAT.

星表交叉融合是天文科研的重要且基础的任务。TOPCAT 软件中提供了多种星表交叉方式，如本地两个星表交叉、本地三个星表的交叉，同时也支持本地星表与 CDS 数据库中的星表交叉。这些功能为天文学家的科研工作提供了便利。

然而，随着天文星表规模的不断增加，当前的数据平台已经无法满足大数据科研的要求，主要体现在两个方面：

第一，随着数据规模的增加，数据下载将耗费大量的时间，成为数据分享使用的主要瓶颈。然而，当前的虚拟天文台平台仍然是以发现数据，查找数据和下载数据为主要的工作模式。因此，在当前进入巡天观测时代，巡天望远镜的星表都在亿级，甚至几十亿级以上，单独星表数据的存储规模都已达到 TB 量级，而对应的测光图像数据，则在百 TB 规模，下载如此规模的数据，需要耗费大量的时间。

第二，在处理亿级规模的星表数据时，传统的科研模式将不再有效。例如，当星表规模达到千万级别时，TOPCAT 加载数据的时间将让人无法接受，即使在高性能的服务器上，TOPCAT 的工作效率也将急剧下降。我们在 12 核 64GB 的虚拟机环境下，使用 TOPCAT 打开一个 2,000 万行的星表，数据加载的时间就超过 30 分钟。其它许多类似的软件也具有相同的问题，数据处理的并行化已经成为当前大数据处理的迫切需求。

6.2　中国虚拟天文台云资源平台

为了改变传统的数据检索、下载、分析的科研模式，我们需要在避免大数据迁移的情况下，实现数据的共享。美国加州理工大学的 G. Bruce Berriman 等 (Bruce et al., 2010) 把商业的云计算应用于天文学，在亚马逊的云计算平台上针对科学

工作流应用来研究数据共享方案。加拿大天文数据中心基于开源的 OpenStack 建立了天文科学平台，实现了数据、存储与计算的协同。这些案例将云计算引入到天文大数据处理上作了有益的偿试。2013 年，中国虚拟天文台（Chinese Virtual Observatory, China-VO) 在中国科学院科技领域云项目的支持下，启动了中国虚拟天文台的天文领域云项目，目标为天文学研究建立了一个混合基础设施环境(Li et al., 2017)，以实现基于大数据的在线科研平台。

6.2.1 平台架构与现状

中国虚拟天文台在线平台是一个网络化的集成科研环境，它最大的特点与优势就是初步实现了一个在线的科研环境原型系统。平台利用云计算技术，通过虚拟化的形式，在数据中心的计算资源池中为天文学家提供按需定制的计算服务，初步实现了如图6-4所示的在线科研形式。

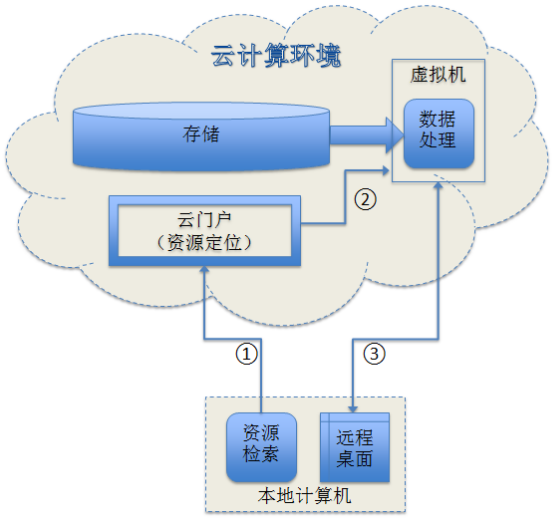


图 6-4 基于中国虚拟天文台平台数据的研究流程。

Figure 6-4 The research process in China-VO platform.

平台在开源云系统 CloudStack[3] 的基础上进行二次开发，整合了统一认证，实现了计算资源的按需定制服务。具有中国科技网通行证的用户登录平台后，可以申请不同配置的计算资源。天文学家个性化的数据处理环境通过模版与快照技术可以随时保存与复制。整个云体系的核心架构如图6-5所示。

平台通过 GlusterFS[4] 分布式文件系统整合不同服务器上的存储资源，形成一个易扩展的存储池。整个存储池挂载在 SAMBA 服务器上，作为 SAMBA 的共享存储系统。利用 SAMBA 的用户授权机制，实现不同用户具有各自不同的独享存储空间。当用户创建自己的科研环境时，虚拟机可以自动挂载属于用户自己的存储空间。平台数据检索系统同样可以将用户检索的数据存储到此存储空间下，从而实现了从平台检索数据到在线处理数据的自动融合。

[3]https://cloudstack.apache.org

[4]https://www.gluster.org

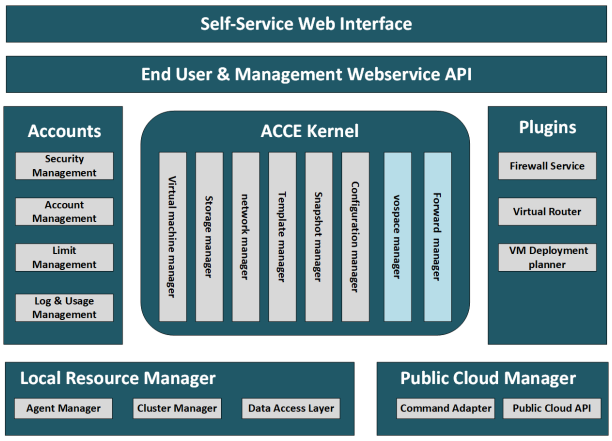


图 6-5 中国虚拟天文台云资源系统体系架构。

Figure 6-5 The architecture of cloud Resource System in China-VO.

然而，由于平台虚拟化计算能力有限，用户创建的虚拟机无法超越单台服务器的能力，导致用户仍然只能在线处理小规模的数据。此外，由于科研环境处在云端，用户需要通过本地电脑连接到虚拟机后才能开展科研工作，如何建立高效稳定的交互式环境也是云资源平台面临的新问题。

6.2.2　高性能计算资源的整合

如果需要在平台上创建大规模数据处理系统，可以有两种方式，一是创建足够多的虚拟机。由于个人存储系统可以在同属于相同帐号下的多台虚拟机间共享，我们可以将这些虚拟机配置成共享存储的集群系统。但这对于天文学家而言，还是感觉有些过于复杂。此外，基于 SAMBA 的共享存储系统的读写性能很难满足大规模并行计算的要求。二是直接在计算集群系统处理，对于大规模计算而言，这样会更加便利。因此，我们考虑如何在云资源架构下整合高性能计算（High performance computing，HPC）系统。

与单个服务器相比，高性能计算系统有三个独特的地方:

(1) 一是计算节点间需要有相同的系统环境，包括操作系统、相关软件库，使得在任意节点上编绎的软件在所有节点上都可执行。

(2) 计算节点间需要有相同的用户系统，从而使得一个用户可以在所有计算节点间相互访问。

(3) 所有计算节点需要同处于一个网段内。

基于高性能计算系统的特殊要求，我们设计了云系统与高性能计算平台的相连架构，如图6-6所示。

云计算与高性能计算整合的关键在于做好计算资源、存储资源、网络资源的规划，使得各组件之间能够互连互通，具体过程可以分成下面两个步骤:

(1) 配置集群登录节点的系统模版，并注册到云资源系统的模版库中，用户可以基于此模版创建虚拟机，此时虚拟机的系统配置是与集群系统的配置是一致的。

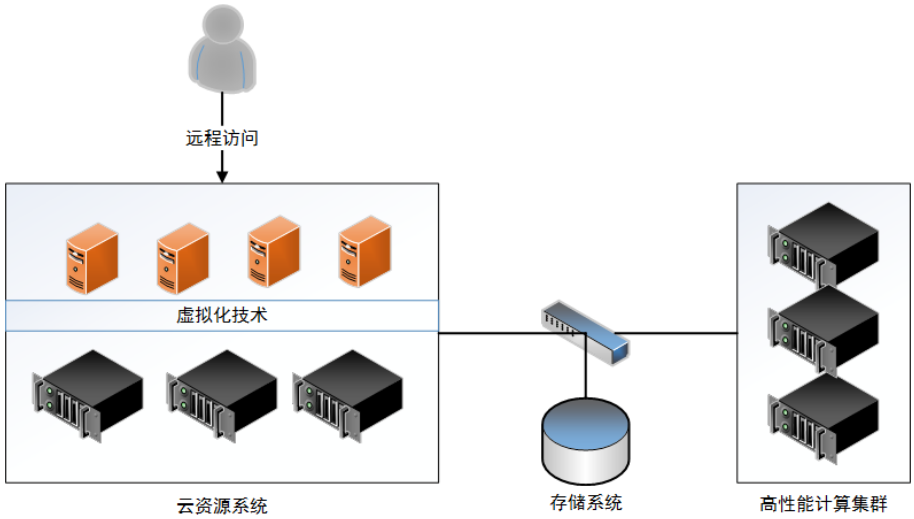


图 6-6 云计算系统与高性能计算系统的整合架构。

Figure 6-6 The integrated architecture between Cloud and HPC.

(2) 在云资源平台创建一个专有分区，并创建私有网络，网络配置与集群的私有网络配置相一致，但需要排除所有物理集群已经使用过的 IP 地址。

通过平台整合，我们可以在云资源平台的专有分区创建虚拟机，而此虚拟机由于网络链路与高性能计算集群相连，环境也相一致，则相当于集群的登录节点。在此环境下，对于小的计算任务，可以直接在虚拟机环境下完成，而如果涉及到大规模的计算任务，我们可以直接提交到集群中执行，从而实现云计算与高性能计算相统一。

6.2.3 基于 MPI 的软件自动并行化

大型巡天数据的处理分析离不开高性能计算环境的支持。高性能计算集群的主要优势就是多计算节点，从而可以将相同的计算任务分发到不同的计算节点上同时运行。高性能计算环境下的并行方式与单机环境明显不同。单机环境下，由于共享内存，通过多进程或多线程方式实现起来比较容易。但是单机系统的计算能力总是比较有限的，而高性能计算环境由于跨计算节点，内存无法共享。当然最笨的办法仍然可以登录到不同的计算节点来启动多个程序，但是当需要大规模并行时，这个办法就无法应对了。比如，需要将 1 亿个天体的波段流量转换为星等，1 亿个天体存储在 100 个 CSV 文件中。如果转换程序一次可以处理一个文件，当串行执行时，那么就需要逐步运行 100 次。而如果同时运行 100 个进程，每个进程处理一个不同的文件，那么程序运行一次的时间就完成了所有天体的星等计算任务，相当于速度提高了 100 倍。

通常，高性能计算环境需要部署 MPI 软件，MPI（Message Passing Interface）是指消息传递接口，定义了高性能计算环境下进程通信的标准接口，具体的实现有很多，比如 MPICH、INTELMPI、OPENMPI 都是比较常用的 MPI 接口库。MPI 程序的开发或修改虽然不是很复杂，但对于天文学家而言，也不是一件容易

的事情。很多软件从串行程序修改为并行程序时，涉及到程序结构、函数等许多代码的调整，或者有可能根本就没有源代码。因此，我们设计了一个代理策略，通过代理来启动数据处理进程，从而在不对程序本身修改的前提下实现了并行执行。

代理策略的核心是代理程序。代理程序本身是一个标准的 MPI 并行程序，支持高性能计算环境下的作业调度系统来启动指定数量的进程。而真正要执行的程序则作为代理程序的参数，代理在进程启动后，在分配的计算节点上执行指定的程序。我们在国家天文科学数据中心集群环境上设计的代理程序名称为 scalempi。我们仍以前面星等转换的程序为例，假设文件名为 1~100, 程序名为 convert.py，如果串行执行，那么依次运行：

```
python convert.py 1
python convert.py 2
...
```

如果采用代理，编写如下作业脚本 convert.pbs：

```
#PBS -N convert
#PBS -l nodes=20:ppn=5
mpirun -genv -machinefile $PBS_NODEFILE -n $NP scalempi 'python convert.py' -r
```

作业提交后，即可同时在集群环境下实现并行。在集群资源充足的情况下，我们可以发现 100 个进程运行在 20 个不同的计算节点上，每个节点运行了 5 个进程，并且每个进程处理不同的文件。

值得注意的是，并行程序虽然运行相同的程序，但每个进程需要处理不同的数据，否则并行再多也没有意义。因此，并行处理时关键问题是需要提前规划好数据，从而实现让不同的进程读取到不同的数据文件。

对于本论文前面的 BASS DR3 数据的分类预测和红移估计工作，采用自动化并行策略，迁移到集群环境下实现，极大地缩短了预测时间。这也充分说明了大数据处理时，并行计算的必要性。

6.2.4　基于集群环境的大规模数据交叉实现

正如第二章所述，天文观测已进入到了全波段时代，包括射电、红外、可见光、紫外、X 射线、直到 γ 射线波段。但是这些波段的数据往往由不同的望远镜来观测，存储在不同数据中心的星表里。如果将这些来自不同波段的异地异构星表联合起来对天体进行研究，就需要星表之间的交叉证认。星表证认主要基于位置，从而得到同一源在不同波段的信息。交叉证认的原理并不复杂，由于位置的误差，不同星表对于同一源的位置也不完全一致，因此通常是采用计算角矩离的方式来判断。当矩离小于给定值时，我们可以认为这两个位置属同一个源。虽然角距离的计算量并不大，但是在星表规模上亿时，完成两两交叉的计算量却是巨大的。目前，两个星表交叉有很多的实现算法，比较常用的就是 CDS 提供的交叉服务，在 TOPCAT 软件中，也提供了相应的交叉功能。

但是，正如在第一节所述，当数据规模不是很大时，比如百万量级，TOPCAT 是深受大家喜爱的天文星表工具。但是数据规模达到千万行时，数据交叉将需要花费大量的时间，当然这也跟服务器的配置直接相关。我们在 32 核、128GB 内存的服务器上采用命令行工具 Stilts 进行了多种不同规模的输入星表的测试，文件 1 为 7GB，400 万行，当文件 2 为 39GB 时，1,600 万行，完成交叉证认花了 1 个小时；当文件 2 再增加一倍时，交叉程序直接报错，无法执行。当然，我们可以增加服务器的配置，把内存做到 TB，这样确实可以增加两表交叉的能力，但是代价也是昂贵的。因此，我们设计了一个将计算迁移到集群环境的方法。

我们设计的方法并不对两表之间的交叉算法进行修改，而只是迁移到高性能计算集群中，从而可以充分利用集群的计算资源，实现不同部分并行执行。基本过程如图6-7所示。假设星表 A 与星表 B 执行交叉，星表 B 为一超大规模星表，我们可以将星表 B 切分成 *n* 个相同大小的文件（B1，B2，...，Bn），将星表 A 分别与星表 B 的子表 B1，B2，...，Bn 进行交叉，然后再将交叉结果全部合并。合并时需要考虑一对多结果的处理方式，如果保留多个交叉结果，则直接合并即可；如果只需要最好的结果，则在合并时对于星表 A 中的任意一个天体只需要保留星表 B 中距离最小的那个对应体即可。

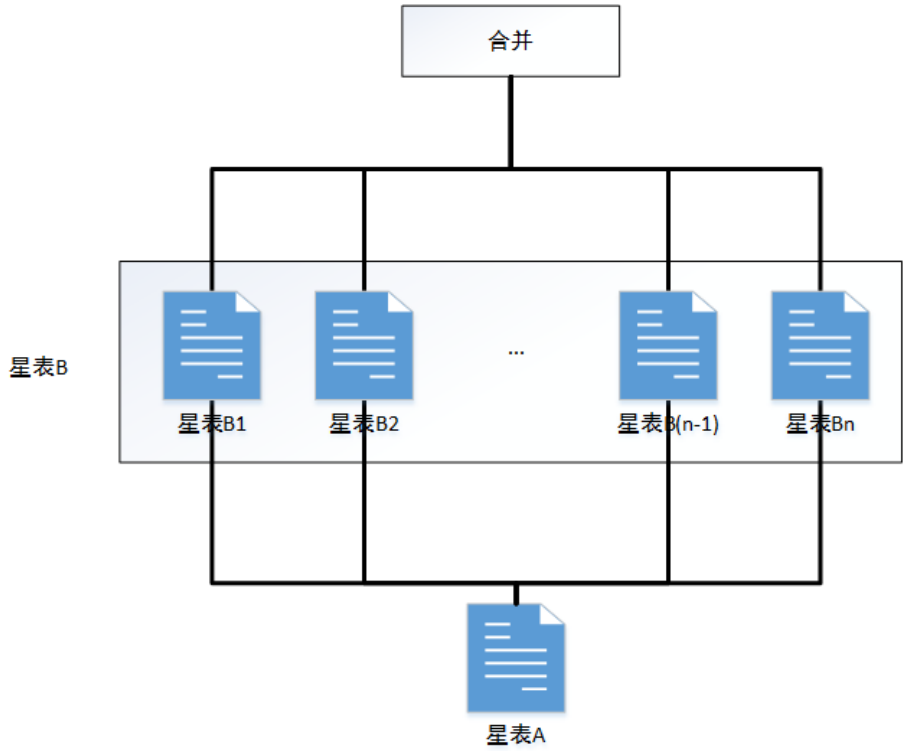


图 6-7 基于集群的并行交叉流程。

Figure 6-7 The parallel cross-match workflow in a HPC.

我们基于上述并行方案，完成了 SDSS DR16 及 LAMOST DR7 的星系分别与 DESI DR9 的测光星表的交叉证认。在具体实施中，DESI DR9 星表被分割成 2,073 个文件，SDSS DR16 的星系分别与 DESI DR9 的 2,073 个文件进行交叉，提

交到集群环境下执行。由于集群资源的限制，我们一次使用 25 个节点，每个节点执行 21 个进程，相当于一次运行 525 个进程，共提交 4 个作业完成全部交叉任务，从而提高了大规模数据之间交叉证认的效率。

6.3　本章小结

本章开始介绍了传统科研模式下数据处理的常用工具及方法。比如，法国斯特拉斯堡数据中心在天文数据、软件与服务等方面走在世界前例。在不涉及超大规模星表的数据处理时，完全可以采用 CDS 的平台或工具完成数据交叉、可视化等科研任务。但是，在大规模数据时代，传统科研模式将导致科研效率低下。中国虚拟天文台在天文大数据的新型科研模式上做了大量的工作。利用云计算技术，设计并实现了可以按需分配的云资源平台，初步实现了在线的数据处理。在此基础上，为了加速类星体选源和测光红移的计算速度，我们设计并开发了基于集群的并行数据交叉系统，极大缩短了大规模星表的交叉证认时间。后续，我们将进一步完善云资源平台与高性能计算集群的整合，降低并行计算的门槛，开发面向类星体选源和测光红移估计的在线计算平台，在此基础上开发更加通用的天文数据计算平台，为大型巡天数据的快速高效处理服务。

第 7 章　结论与展望

7.1　研究结论

我们主要利用大型测光巡天数据开展了类星体候选体的选源方法及类星体、星系的测光红移方法的研究。同时为提高大规模天文数据处理的效率而开展了在线科研平台及计算并行化等技术攻关和算法优化。数据量大是大型巡天观测时代的主要特征之一。因此，对于采用机器学习方法进行模型构建不仅要考虑准确性，同时也需要考虑训练与预测的时间。经过调研与对比，对于本论文采用的已知样本，以 CatBoost、XGBoost 为代表的 'Booting' 方式的集成学习算法在准确度、训练时间等方面都具有较大的优势，适合于天体分类与红移预测的回归任务。将构建的分类与回归模型应用于巡天数据，我们共发布了三个星表：BASS DR3 天体分类星表、BASS DR3 类星体候选体的红移估测星表及 DESI DR9 星系的红移估测星表，为后续进一步的研究提供了分类基础和物理参数测量。

7.1.1　类星体选源

BASS DR3 是我国目前自产的最大规模图像测光数据集。我们采用 SDSS 与 LAMOST 巡天的光谱数据与 BASS DR3 数据交叉，获得了较大的已知样本。再将样本与 ALLWISE 交叉，得到对应天体的红外特征 ($W1$、$W2$)。我们对已知样本在星等-颜色空间上进行了分析。采用三个形态学特征，即 g、r、z 波段的 Kron 星等与 PSF 星等之差，对于展源（星系）与点源（恒星和类星体）分类的准确度可以达到 90%。但这些特征无法进行恒星与类星体的分类。综合考虑可见光和红外波段特征，以及已知光谱信息，我们应用 XGBoost 分类算法构建天体的分类模型。我们采用了两种分类策略：第一种策略为采用多元分类模型，在已知星系、恒星及类星体的样本上构建能够提供星系、恒星及类星体三种分类标签的分类模型，实验结果表明最优分类特征为 $z-W2$, Δz, $W1-W2$, Δr, $g-r$, $z-W1$, Δg, $g-z$, $r-W2$, $r-z$, r，得到的总分类准确率为 98.43%；第二种策略采用分层二元分类模型，即先进行点源与展源的分类，然后再将点源分为恒星与类星体。点源与展源分类模型采用的最优输入特征为 Δg, Δz, $g-W1$, $W1-W2$, $z-W1$, Δr, $g-z$, $z-W2$, $g-r$,$r-z$,$W1$, r, g, z, $r-W2$，总准确率为 98.67%；恒星与类星体分类模型采用的最优输入特征为 $z-W2$, $W1-W2$, $g-z$, $g-r$, $z-W1$, $r-z$, Δz, r, $r-W2$, z, Δg, $g-W1$, Δr, $W1$, $g-W2$，总准确率为 99.15%。考虑到 BASS DR3 中大部分天体并没有红外对应体，因此，我们也训练了只采用光学特征的分类模型。利用两种策略训练的分类模型完成了对 BASS DR3 数据的分类。如果考虑高可靠性，我们选择同时具有光学和红外特征，两种分类策略下都被分为类星体且可能性概率超过 95% 的天体作为类星体候选体，共有 798,928 个源。我们将这些候选体与 SIMBAD 以 1 角秒为半径进行交叉证认，获得 184,376 个匹配源，其中 175,292 个源是类星体，即 95% 的天体已经被证认为是类星体，从

而证明了本研究方法的有效性。

7.1.2 测光红移估计

7.1.2.1 类星体的红移估计

在 BASS DR3 数据分类的基础上，我们同时采用 CatBoost、XGBoost 与随机森林方法对类星体候选体进行了红移估计。我们采用 SDSS DR16Q 与 LAMOST DR5 中的类星体星表与 BASS DR3 交叉得到已知样本。已知样本的红移覆盖了较宽的范围，我们采用两种方案分别进行了红移估计。第一种方案称为一步模型，采用 CatBoost 方法，将已知样本作为整体进行训练，最优模型时的最佳输入特征为 $\Delta z, z-W2, g-r, W1, r-z, z-W1, W1-W2, r-W2, \Delta g, g-z, z, W2, \Delta r$，此时构建的模型的 MSE 为 0.1600。第二种方案称为两步模型，即将已知样本以 3.5 为界按高、低红移进行分类，然后再对高红移、低红移样本分别进行了红移估计。对于高红移样本，最优性能 MSE 为 0.0756；而对于低红移样本，MSE 为 0.1497。由于高红移与低红移样本的不平衡性，一步模型下容易将高红移低估，而两步模型，对于高红移样本的预测性能有了明显的提升。我们使用两种方案对 BASS DR3 数据中的所有类星体候选体进行了红移预测。

7.1.2.2 星系的红移估计

相比于类星体，普通星系具有更大的样本空间。我们采用 SDSS DR16、LAMOST DR7 光谱数据与 DESI DR9 进行交叉，构建已知星系样本。通过已知样本与待测样本的对比，已知样本明显偏亮，红移最高不超过 2。因此，对于亮源（r 星等在 22 等以内），训练样本充足，而对于暗源，则明显短缺。因此，我们采用了机器学习与模板匹配相结合的方法来进行红移估计。利用 EAZY 的模板匹配方法及软件工具，我们完成了 DESI DR9 中所有星系的红移预测。对于已知样本中的星系，在预测质量参数 $Q_z < 1$ 时，对已知样本的预测性能为 $MSE = 0.03$。同时，采用 CatBoost 方法进行回归训练，结合光学、红外及孔颈特征，构建的回归模型性能为 $MSE = 0.0032$。结果表明，在已知样本充足的情况下，机器学习算法具有明显的优势。而对于暗源的红移预测，模板匹配方法不失为有益的补充。

7.1.3 面向类星体选源与红移估计的在线科研平台

大型巡天时代所产生的大规模数据对于传统天文学研究方法提出了巨大的挑战，包括对数据的计算、存储及传输。尤其是传输，成为了大数据科研的最大瓶颈，数据无法随时复制到本地机器上。面向类星体选源与红移估计的在线计算平台在结合国家天文科学数据中心现有云计算平台的基础上，实现了集群计算资源的整合。对于数据准备、模型应用两个阶段的大数据处理分析设计了并行算法，提高了大数据的处理效果。通过在线计算平台的应用与实践，为天文大数据科研平台的设计打下基础。

7.2　研究创新点

(1) 利用成熟的集成机器学习算法与大型巡天数据相结合，通过设计多种应用策略，在大规模已知样本上分类和回归取得了时间与精度的双提升，不仅对全部的 BASS 巡天数据给出分类，其中类星体候选体可以作为 LAMOST 输入星表，而且还发现了一大批高红移类星体候选体源表。通过本研究的多个实验，形成了定位最优特征及模型参数的方法，为后续其它巡天数据的分类提供了借鉴。

(2) 采用传统模板匹配与机器学习相结合的方法，给出了 DESI 巡天的星系测光红移的估测星表，为新型的大型图像巡天数据准确快速地估计星系红移提供了解决方案。这些方法将来也可以应用于 LSST、CSST 等巡天项目的星系红移预测。

(3) 通过构建与完善面向类星体选源与及测光红移估计的在线科研平台，为天文学家提供远程开展大数据分析的全新工作模式，推动大数据时代天文学研究范式的转变。同时，通过设计软件自动并行化方法，实现了大规模数据处理的并行化，提高科研效率。

7.3　后续展望

7.3.1　开发基于深度学习的选源与红移测量方法

传统的机器学习方法的输入特征都是已经处理过的数据，特征数量有限，且提取的特征值与测光软件紧密相关。比如 BASS DR3 与 DESI DR9 的北天部分，虽然采用了相同的图像，但发布的测光星表确有显著的差异。而深度学习方法可以从原始图像直接进行学习，更容易发现细微的变化。深度学习是人工智能与机器学习领域的前沿研究方向，我们可以利用已经训练好的深度网络模型，针对不同巡天数据与学习任务进行适应调整，从而构建适合新的模型。

7.3.2　结合多个不同巡天数据进行选源与红移测量方法优化

本研究采用的巡天数据只包括了 g、r、z 三个光学波段，但只依赖于这三个光学波段，很难区分出类星体与恒星。因此，我们结合了 ALLWISE 的中红外波段数据，使得分类结果有了很大提升。此外，类星体在紫外、射电、X 射线、γ 射线等许多波段都具有显著特征。我们可以结合更多的波段数据进一步提升类星体选源的准确率。同样的，更多的波段数据也能改善红移估计的准确性。

7.3.3　进一步发展模板匹配与机器学习相结合的红移估计方法

随着多个图像巡天望远镜即将运行，大规模更暗的天体将被发现。而光谱巡天的效率较低，短时间内很难充实暗弱天体的光谱红移样本。因此，基于模板匹配与机器学习相结合的方法是解决这一问题的有效手段。这一方法除了发现更高效的机器学习算法外，另一个重要的任务是要充实模板库，从而提高暗源星系的红移估计的准确率。

7.3.4 针对 CSST 项目开展天体分类和红移测量方法研究

调研和研究中国巡天空间望远镜（CSST）项目的数据特点，基于 CSST 的模拟数据开展分类和回归算法研究。同时，收集和丰富已知样本，在与 CSST 相似的真实数据上，挑选适合 CSST 数据的最优方法。

参考文献

Abazajian K, AdelmanMcCarthy J K, Agueros M A, et al. The first data release of the sloan digital sky survey[J]. The Astronomical Journal, 2003, 126(4): 2081-2086.

Abazajian K, AdelmanMcCarthy J K, Agueros M A, et al. The second data release of the sloan digital sky survey[J]. The Astronomical Journal, 2004, 128(1): 502-512.

Abazajian K, AdelmanMcCarthy J K, Agueros M A, et al. The third data release of the sloan digital sky survey[J]. The Astronomical Journal, 2005, 129(3): 1755-1759.

Abazajian K N, AdelmanMcCarthy J K, Agueros M A, et al. The seventh data release of the sloan digital sky survey[J]. The Astrophysical Journal Supplement, 2009, 182(2): 543-558.

Abdurro'uf, Katherine A, Conny A, et al. The seventeenth data release of the sloan digital sky surveys: Complete release of manga, mastar, and apogee-2 data[J]. The Astrophysical Journal Supplement Series, 2022, 259(2): 39.

Abolfathi B, Aguado D S, Aguilar G, et al. The fourteenth data release of the sloan digital sky survey: First spectroscopic data from the extended baryon oscillation spectroscopic survey and from the second phase of the apache point observatory galactic evolution experiment[J]. The Astrophysical Journal Supplement Series, 2018, 235(2).

Adelman-McCarthy J K, Agueros M A, Allam S S, et al. The fouth data release of the sloan digital sky survey[J]. The Astrophysical Journal Supplement, 2006, 162(1): 38-48.

Adelman-McCarthy J K, Agueros M A, Allam S S, et al. The fifth data release of the sloan digital sky survey[J]. The Astrophysical Journal Supplement, 2007, 172(2): 634-644.

Adelman-McCarthy J K, Agueros M A, Allam S S, et al. The sixth data release of the sloan digital sky survey[J]. The Astrophysical Journal Supplement, 2008, 175(2): 297-313.

Aguado D S, Ahumada R, Almeida A, et al. The fifteenth data release of the sloan digital sky surveys: First release of manga-derived quantities, data visualization tools, and stellar library[J]. The Astrophysical Journal Supplement Series, 2019, 240(2).

Ahn C P, Alexandroff R, Allende-Prieto C, et al. The ninth data release of the sloan digital sky survey: First spectroscopic data from the sdss-iii baryon oscillation spectroscopic survey[J]. The Astrophysical Journal Supplement, 2012, 203(2).

Ahn C P, Alexandroff R, Allende-Prieto C, et al. The tenth data release of the sloan digital sky survey: First spectroscopic data from the sdss-iii apache point observatory galactic evolution experiment [J]. The Astrophysical Journal Supplement, 2014, 211(2).

Ahumada R, Prieto C A, Almeida A, et al. The 16th data release of sloan digital sky survey: First release from the apogee-2 southern survey and full release of eboss spectra[J]. The Astrophysical Journal Supplement Series, 2020, 249(3): 21.

Aihara H, Allende Prieto C, An D, et al. The eighth data release of the sloan digital sky survey: First data from sdss-iii[J]. The Astrophysical Journal Supplement, 2011, 193(2).

Alam S, Albareti F D, Allende-Prieto C, et al. The eleventh and twelfth data releases of the sloan digital sky survey: Final data from sdss-iii[J]. The Astrophysical Journal Supplement, 2015, 219 (1).

Albareti F D, Allende-Prieto C, Almeida A, et al. The 13th data release of the sloan digital sky survey: First spectroscopic data from the sdss-iv survey mapping nearby galaxies at apache point observatory[J]. The Astrophysical Journal Supplement Series, 2017, 233(2).

Arnason R M, Barmby P, Vulic N. Identifying new x-ray binary candidates in m31 using random forest classification[J]. Monthly Notices of the Royal Astronomical Society, 2020, 492(4): 5075-5088.

Arnouts S, Cristiani S, Moscardini L, et al. Measuring and modelling the redshift evolution of clustering: the hubble deep field north[J]. Monthly Notices of the Royal Astronomical Society, 1999, 310(2): 540-556.

Arviset C, Gaudet S, the IVOA Technical Coordination Group. https://ivoa.net/documents/notes/ivoaarchitecture/20101123/ivoaarchitecture-1.0-20101123.pdf [M]. On-line Resources, 2010.

Babbedge T S R, Rowan-Robinson M, Gonzalez-Solares E, et al. Impz: a new photometric redshift code for galaxies and quasars[J]. Monthly Notices of the Royal Astronomical Society, 2004, 353 (2): 654-672.

Ball N M, Brunner R J, Myers A D, et al. Robust machine learning applied to astronomical data sets. ii. quantifying photometric redshifts for quasars using instance-based learning[J]. The Astrophysical Journal, 2007, 663(2): 774-780.

Baum W A. Photoelectric determinations of redshifts beyond 0.2c[J]. Astrophysical Journal, 1957, 62(1): 6-7.

Baum W A. Photoelectric magnitudes and red-shifts[J]. Proceedings from IAU Symposium no. 15, Macmillan Press, New York, 1962.

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): 227-262.

Benitez N. Bayesian photometric redshift estimation[J]. The astrophysical journal, 2000, 536(2): 571-583.

Bethapudi S, Desai S. Separation of pulsar signals from noise using supervised machine learning algorithms[J]. Astronomy and Computing, 2018, 23(1): 15-26.

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 Astrophysical Journal, 2017, 154(1): 28.

Bolzonella M, Miralles J, Pellô R. Photometric redshifts based on standard sed fitting procedures [J]. Astronomy and Astrophysics, 2000, 363: 476-492.

Bonfield D G, Sun Y, Davey N, et al. Photometric redshift estimation using gaussian processes[J]. Monthly Notices of the Royal Astronomical Society, 2010, 405(2): 987-994.

Bovy J, Myers A D, Hennawi J F, et al. photometric redshifts and quasar probabilities from a single, data-driven generative model[J]. The Astrophysical Journal, 2012, 749: 20.

Boyle B J, Shanks T, Croom S M, et al. The 2df qso redshift survey - i. the optical luminosity function of quasi-stellar objects[J]. Monthly Notices of the Royal Astronomical Society, 2000, 317(4): 1014-1022.

Brad N H, W. L.and Alexandra, McLane J N, S. D P. The sloan digital sky survey quasar catalog: Sixteenth data release[J]. The Astrophysical Journal Supplement Series, 2020, 250(8): 24.

Brammer G B, Van Dokkum P G, Coppi P. Eazy: A fast, public photometric redshift code[J]. The astrophysical journal, 2008, 686: 1503-1513.

Breiman L. Random forests[J]. Machine Learning, 2001, 45: 5-32.

Bruce G, Deelman E, Groth P, et al. The application of cloud computing to the creation of image mosaics and management of their provenance: 1006.4860[A]. 2010.

Cai Z, Zhang C, Fan F. History and prospect of astronomical telescopes: Introducing tsinghua multiplexed survey telescope (MUST)[J]. Experimental Technology and Management, 2021, 38 (5): 1-9.

Carliles S, Budavári T, Heinis S, et al. Random forests for photometric redshifts[J]. The Astrophysical Journal, 2010, 712(1): 511-515.

Chawla N V, Bowyer K W, Hall L O, et al. Smote: Synthetic minority over-sampling technique[J]. Journal Of Artificial Intelligence Research, 2002, 16(1): 321-357.

Chen T, Guestrin C. Xgboost: A scalable tree boosting system[A]. 2016.

Clarke A O, Scaife A M M, Greenhalgh R, et al. Identifying galaxies, quasars, and stars with machine learning: A new catalogue of classifications for 111 million sdss sources without spectra [J]. Astronomy & Astrophysics, 2020, 639.

Coronado-Bl*â*zquez J. Classification of fermi-lat unidentified gamma-ray sources using catboost gradient boosting decision trees[J]. Monthly Notices of the Royal Astronomical Society, 2022, 515(2): 1807-1814.

Cui C, Ye C, Xiao J, et al. Astronomy research in big-data era[J]. Chinese Science Bulletin, 2015, 60(5-6): 445-449.

Cui X Q, Zhao Y H, Chu Y Q, et al. The large sky area multi-object fiber spectroscopic telescope (LAMOST)[J]. Research in Astronomy and Astrophysics, 2012, 12(9): 1197-1242.

Curran S J, Moss J P, Perrott Y C. Qso photometric redshifts using machine learning and neural networks[J]. Monthly Notices of the Royal Astronomical Society, 2021, 503(1): 11.

Das P, Sanders J. Made: a spectroscopic mass,age,and distance estimator for red giant stars with bayesian machine learning[J]. Monthly Notices of the Royal Astronomical Society, 2019, 484(1): 294-304.

Dey A, Schlegel D J, Lang D, et al. Overview of the desi legacy imaging surveys[J]. The Astronomical Journal, 2019, 157(5): 29.

DiPompeo M A, Bovy J, D. M A, et al. Quasar probabilities and redshifts from wise mid-ir through galex uv photometry[J]. Monthly Notices of the Royal Astronomical Society, 2015, 452: 3124-3138.

Dorogush A v, Ershov V, G. Y A. Catboost: gradient boosting with categorical features support: arXiv:1810.11363[Z]. 2018.

Euclid Collaboration. Euclid preparation x. the euclid photometric-redshift challenge[J]. Astronomy & Astrophysics, 2020, 644.

Fabbro S, Venn K A, O'Briain T, et al. An application of deep learning in the analysis of stellar spectra[J]. Monthly Notices of the Royal Astronomical Society, 2018, 475(1): 2978-2993.

Fadely R, Hogg D W, Willman B. Star-galaxy classification in multi-band optical imaging[J]. The Astrophysical Journal, 2012, 760(1): 10.

Feldmann R, Carollo C M, Porciani C, et al. The zurich extragalactic bayesian redshift analyzer and its first application: Cosmos[J]. Monthly Notices of the Royal Astronomical Society, 2006, 372 (2): 565-577.

Firth A, Lahav O, Somerville R. Estimating photometric redshifts with artificial neural networks[J]. Monthly Notices of the Royal Astronomical Society, 2003, 339(4): 1195-1202.

Foltz C B, Chaffee F H, Hewett P C, et al. The apm qso survey. i. initial mmt result[J]. Astronomical Journal, 1987, 94(1): 1423-1460.

Fu Y, Wu X B, Yang Q, et al. Finding quasars behind the galactic plane. i. candidate selections with transfer learning[J]. The Astrophysical Journal Supplement Series, 2021, 254(6): 20.

Gao D, Zhang Y, Zhao Y. Support vector machines and kd-tree for separating quasars from large survey data bases[J]. Monthly Notices of the Royal Astronomical Society, 2008, 386(3): 1417-1425.

Gregg M D, Becker R H, White R L, et al. The first bright qso survey[J]. Astronomical Journal, 1996, 112: 407-426.

Han J, Kamber M. Data mining: Concepts and techniques[M]. Singapore: Elsevier, 2006.

Hawkins M R S. Variable extragalactic objects: identification and analysis of a complete sample to b =21[J]. Monthly Notices of the Royal Astronomical Society, 1983, 202: 571-585.

He X T, Wu J H, Yuan Q R, et al. The multiwavelength quasar survey. i. initial results[J]. The Astrophysical Journal, 2001, 121(4): 1863-1871.

Henghes B, Pettitt C, Thiyagalingam J, et al. Benchmarking and scalability of machine-learning methods for photometric redshift estimation[J]. Monthly Notices of the Royal Astronomical Society, 2021, 505(1): 19.

H*â*la P. Spectral classification using convolutional neural networks[A]. 2014.

Hickox R C, Jones C, Forman W R, et al. A large population of mid-infrared-selected, obscured active galaxies in the bootes field[J]. The Astrophysical Journal, 2007, 671: 1365-1387.

Huertas-Company M, Rouan D, Tasca L, et al. A robust morphological classification of high-redshift galaxies using support vector machines on seeing limited images. i.method description[J]. Astronomy and Astrophysics, 2008, 478(3): 971-980.

Ishida E E O, Beck R, Gonz*â*lez-Gait*â*n S, et al. Optimizing spectroscopic follow up strategies for supernova photometric classification with active learning[J]. Monthly Notices of the Royal Astronomical Society, 2019, 483(1): 2-18.

Jerome H F. Greedy function approximation: A gradient boosting machine[J]. The Annals of Statistics, 2001, 29(5): 1189-1232.

Jiang L, Ning Y, Fan X, et al. Definitive upper bound on the negligible contribution of quasars to cosmic reionization[J]. nature astronomy, 2022, 6: 850-856.

Jiang P, Yue Y, Gan H, et al. Commissioning progress of the fast[J]. Science China physics, mechanics & Astronomy, 2019, 62(5).

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(1): 4539-4549.

Kind M C, Brunner R J. Somz: photometric redshift pdfs with self-organizing maps and random atlas[J]. Monthly Notices of the Royal Astronomical Society, 2014, 438: 3409-3421.

Kleinmann S G. Robotic telescopes in the 1990s,103rd Annual meeting of the astronomical society of the pacific, Univ. of Wyoming[C]. Laramie, 1992.

Kormendy J, Ho L C. Coevolution (or not) of supermassive black holes and host galaxies[J]. Annual Review of Astronomy and Astrophysics, 2013, 51(1): 511-653.

Krakowski T, Malek K, Bilicki M, et al. Machine-learning identification of galaxies in the wisexsupercosmos all-sky catalogue[J]. Astronomy & Astrophysics, 2016, 596(1): 11.

Ksoll V F, Gouliermis D A, Klessen R S, et al. Hubble tarantula treasury project-vi. identification of pre-main-sequence stars using machine-learning techniques[J]. Monthly Notices of the Royal Astronomical Society, 2018, 479(2): 2389-2414.

Le-Borgne D, Rocca-Volmerange B. Photometric redshifts from evolutionary synthesis with p$\hat{E}$gase: The code z-peg and the z=0 age constraint[J]. Astronomy and Astrophysics, 2002, 386: 446-455.

Li C, Cui C, Mi L, et al. Design and implement of astronomical cloud computing environment in china-vo[J]. proceedings IAU Symposium No. 325, Astroinformatics, 2017.

Loh E D, Spillar D J. Photometric redshifts of galaxies[J]. Astrophysical Journal, 1986, 303(1): L54-L61.

LSST Science Collaborations. LSST Science Book, Version 2.0[M]. ArXiv:0912.0201, 2009.

Luo A L, Zhao Y H, Zhao G, et al. The first data release (dr1) of the lamost regular survey[J]. Research in Astronomy and Astrophysics, 2015, 15(8): 1095-1124.

Malek K, Solarz A, Pollo A, et al. The vimos public extragalactic redshift survey (vipers). a support vector machine classification of galaxies, stars, and agns[J]. Astronomy and Astrophysics, 2013, 557(1): 16.

Marocco F, Eisenhardt P R M, Fowler J W, et al. The catwise2020 catalog[J]. The Astrophysical Journal Supplement Series, 2021, 253(1): 22.

Meusinger H, Scholz R D, Irwin M, et al. Qsos from the variability and proper motion survey in the m 3 field[J]. Astronomy and Astrophysics, 2002, 392: 851-863.

Mirabal N, Charles E, Ferrara E, 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): 8.

Mitchell T M. Machine learning, international edition[M]. McGraw-Hill, 1997.

Palanque-Delabrouille N, Magneville C, Yeche C, et al. The extended baryon oscillation spectroscopic survey:variability selection and quasar luminosity function[J]. Astronomy & Astrophysics, 2016, 587(1): A41.

Parks D, Prochaska J X, Dong S, et al. Deep learning of quasar spectra to discover and characterize damped lyα systems[J]. Monthly Notices of the Royal Astronomical Society, 2018, 476(1): 1151-1168.

Pasquet-Itam J, Pasquet J. Deep learning approach for classifying, detecting and predicting photometric redshifts of quasars in the sloan digital sky survey stripe 82[J]. Astronomy & Astrophysics, 2018, 611(2): A97.

Pearson K A, Palafox L, Griffith C. Searching for exoplanets using artificial intelligence[J]. Monthly Notices of the Royal Astronomical Society, 2018, 474(1): 478-491.

Peng N, Zhang Y, Zhao Y, et al. Selecting quasar candidates using a support vector machine classification system[J]. Monthly Notices of the Royal Astronomical Society, 2012, 425(4): 2599-2609.

Pichara K, Protopapas P, Kim D W, et al. An improved quasar detection method in eros-2 and macho lmc data sets[J]. Monthly Notices of the Royal Astronomical Society, 2012, 427(1): 1284-1297.

Plewa P M. Random forest classification of stars in the galactic centre[J]. Monthly Notices of the Royal Astronomical Society, 2018, 476(3): 3974-3980.

Puschell J J, Owen F N, Laing R A. Near-infrared photometry of distant radio galaxies-spectral flux distributions and redshifts estimates[J]. Astrophysical Journal, 1982, 257(1): L57-L61.

Salvato M, llbert O, Hoyle B. The many flavours of photometric redshifts[J]. nature astronomy, 2019, 3: 212-222.

Schindler J, Fan X, McGreer I, et al. The extremely luminous quasar survey in the sdss footprint. i. infrared-based candidate selection[J]. ApJ, 2007, 851(1): 13.

Schindler J T, Fan X H, McGreer I D, et al. The extremely luminous quasar survey in the sdss footprint. i. infraredbased candidate selection[J]. The Astrophysical Journal, 2017, 851(1): 8.

Schlegel D J, Finkbeiner D P, Davis M. Maps of dust infrared emission for use in estimation of reddening and cosmic microwave background radiation foregrounds[J]. The Astrophysical Journal, 1998, 500(2): 525-553.

Schmidt M. 3c 273: A star-like object with large red-shift[J]. Nature, 1963, 197.

Schmidt M, Green R F. Quasar evolution derived from the palomar bright quasar survey and other complete quasar surveys[J]. Astrophysical Journal, 1983, 269(1): 352-374.

Schmidt S J, Malz A I, Soo J Y H, et al. Evaluation of probabilistic photometric redshift estimation approaches for the rubin observatory legacy survey of space and time (LSST)[J]. Monthly Notices of the Royal Astronomical Society, 2020, 499: 1587-1606.

Shallue A, C. J.and Vanderburg. Identifying exoplanets with deep learning: A five-planet resonant chain around kepler-80 and an eighth planet around kepler-90[J]. The Astronomical Journal, 2018, 155(2): 21.

Singh H, Gulati R, Gupta R. Stellar spectral classification using principal component analysis and artificial neural networks[J]. Monthly Notices of the Royal Astronomical Society, 1998, 295(1): 312-318.

Snider S, Prieto C, Hippel T, et al. Three-dimensional spectral classification of low-metallicity stars using artificial neural networks[J]. The Astrophysical Journal, 2001, 562(1): 528-548.

Soifer B T, Sanders D B, Neugebauer G, et al. The luminosity function and space density of the most luminous galaxies in the iras survey[J]. Astrophysical Journal Letters, 1986, 303: L41.

Stern D, Eisenhardt P, Gorjian V, et al. Mid-infrared selection of active galaxies[J]. The Astrophysical Journal, 2005, 631: 163-168.

Storrie-Lombardi M C, Lahav O, Sodre L J, et al. Morphological classification of galaxies by artificial neural networks[J]. Monthly Notices of the Royal Astronomical Society, 1992, 259(1): 8-12.

Stoughton C, Lupton R H, Bernardi M, et al. Sloan digital sky survey: Early data release[J]. The Astronomical Journal, 2002, 123(1): 485-548.

Tagliaferri R, Longo G, Andreon S, et al. Neural networks for photometric redshifts evaluation[J]. Neural Nets. Lecture Notes in Computer Science, 2003, 2859(1): 226-234.

Teimoorinia H, Bluck A F L, Ellison S L. An artificial neural network approach for ranking quenching parameters in central galaxies[J]. Monthly Notices of the Royal Astronomical Society, 2016, 457(1): 2086-2106.

Thompson D J, Bertsch D L, Dingus B L, et al. The second egret catalog of high-energy gamma-ray sources[J]. The Astrophysical Journal Supplement Series, 1995, 101: 259.

Tyson J A. Large synoptic survey telescope: Overview[J]. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 2002, 4836(1): 10-20.

Vanzella E, Cristiani S, Fontana A, et al. Photometric redshifts with the multilayer perceptron neural network: Application to the hdf-s and sdss[J]. Astronomy and Astrophysics, 2004, 423(1): 761-776.

Vapnik V N. The nature of statistical learning theory[M]. New York: Springer, 1995.

Wang D, Zhang Y X, Liu C, et al. Kernel regression for determining photometric redshifts from sloan broad-band photometry[J]. Monthly Notices of the Royal Astronomical Society, 2007, 382 (4): 1601-1606.

Way M J, Klose C D. Can self-organizing maps accurately predict photometric redshifts[J]. Publications of the astronomical society of the pacific, 2012, 124(913): 274-279.

Way M J, Srivastava A N. Novel methods for predicting photometric redshifts from broadband photometry using virtual sensors[J]. The astrophysical journal, 2006, 647: 102-115.

Way M J, Foster L V, Gazis P R, et al. New approaches to photometric redshift prediction via gaussian process regression in the sloan digital sky survey[J]. The astrophysical journal, 2009, 706: 623-636.

White R L, Becker R H, Gregg M D, et al. The first bright quasar survey. ii. 60 nights and 1200 spectra later[J]. The Astrophysical Journal Supplement Series, 2000, 126(2): 133-207.

Witten I H, E. F. Data mining: Practical machine learning tools and techniques[M]. San Francisco:Morgan Kaufmann Publishers, 2011.

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 (1): 1868-1881.

Wu X B, Zhang W, Zhou X. Color-redshift relations and photometric redshift estimations of quasars in large sky surveys[J]. Chinese Journal of Astronomy & Astrophysics, 2004, 4(1): 17-27.

Wu X B, Zuo W, Yang J, et al. Discovering bright quasars at intermediate redshifts based on optical/near-infrared colors[J]. The Astronomical Journal, 2013, 146(4).

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.

Wu X B, Hao G, Jia Z, et al. Sdss quasars in the wise preliminary data release and quasar candidate selection with optical/infrared colors[J]. The Astronomical Journal, 2012, 144(2).

Xiang G, Chen J, Qiu B, et al. Estimating stellar atmospheric parameters from the lamost dr6 spectra with scdd model[J]. Publications of the Astronomical Society of the Pacific, 2021, 133(1): 12.

Yang Q, Wu X B, Fan X H, et al. Quasar photometric redshifts and candidate selection: A new algorithm based on optical and midinfrared photometric data[J]. The Astrophysical Journal, 2017, 154(6): 269.

Zhang Y, Zhao Y. Automated clustering algorithms for classification of astronomical objects[J]. Astronomy & Astrophysics, 2004, 422(3): 1113-1121.

Zhang Y, Ma H, Peng N, et al. Estimating photometric redshifts of quasars via the k-nearest neighbor approach based on large survey databases[J]. The Astronomical Journal, 2013, 146(2).

Zou H, Fan X, Zhang T, et al. Project overview of the beijing-arizona sky survey[J]. Publications of the Astronomical Society of the Pacific, 2017a, 129(1): 9.

Zou H, Zhang T, Zhou Z, et al. The first data release of the beijing-arizona sky survey[J]. The Astronomical Journal, 2017b, 153(1): 14.

Zou H, Zhang T, Zhou Z, et al. The second data release of the beijing-arizona sky survey[J]. The Astrophysical Journal Supplement Series, 2018, 237(1): 15.

Zou H, Zhou X, Fan X, et al. The third data release of the beijing-arizona sky survey[J]. The Astrophysical Journal Supplement Series, 2019, 245(1): 17.

致 谢

2020 年，新冠疫情席卷全球。博士入学考试首次在南京天光所举行，当时的情景还历历在目。已是不惑之年的我，心中即兴奋又忐忑。感谢幸运之神的眷顾，让我顺利开启了这一段博士之旅。在此，感谢国家天文台所给予的学习机会。

时间总是悄悄地流逝。不经意间，博士的学习旅程已接近尾声。三年的博士学习生活是紧张而忙碌的，充满了挑战。博士学习让我真正走进了天文学的世界，领略了宇宙的博大，激发了我探索宇宙的热情。

博士学业能够顺利完成，首先感谢崔辰州研究员。正是在崔老师的鼓励与建议下，才让我有信心重拾学业。崔老师知识渊博，涉猎广泛，具有敏锐的洞察力，让我敬佩。崔老师对科研方向的执著和脚踏实地的作风，20 年来带领中国虚拟天文台跨过一个又一个台阶，走到了国际虚拟天文台的领域前沿。在课题研究中，崔老师给了我充分的自由度，对科研方法及方向给了我许多具体的建议。

感谢国家天文台张彦霞研究员。张老师具有扎实的天文理论基础和丰富的实践经验。从论文选题到各项科研工作的开展都受益于张老师的及时全面的指导，引导我真正进入天文学的殿堂。博士期间撰写的每篇论文都凝聚了她的智慧与汗水。张老师对科学的严谨态度，对论文逐字逐句仔细推敲，数据必须准确无误的要求是我学习的榜样。

感谢广州大学的王锋教授、华南师范大学的李乡儒教授、北京师范大学的胡彬教授、紫金山天文台的左喜营研究员，新疆天文台的王娜研究员，张海龙研究员、国家天文台的姜鹏研究员、彭勃研究员、张海燕研究员、李金增研究员、王启明研究员、潘高峰研究员对我的选题及进展提出的许多有益建议。

感谢北京大学吴学兵教授、广州大学王锋教授、华南师范大学的李乡儒教授、国家天文台赵永恒研究员、陆由俊研究员、邹虎研究员在百忙之中参加我的博士论文答辩，并对论文和未来科研工作提出的诸多建议。

感谢教育处的梁艳春老师、杜红荣老师、马怀宇老师、李响老师，正是因为他们无微不至的工作，才使得我在博士期间的学习与科研能够顺利开展。

感谢国家天文台赵永恒研究员、高亮研究员、施建荣研究员、罗阿理研究员、刘超研究员、赵有研究员、王杰研究员、邹虎研究员、吴潮研究员在学习上所给予的指导。

感谢天津大学的于策博士、肖健博士给予的帮助，每一次的交流与讨论都让我受益非浅。

感谢同项目组的何勃亮、樊东卫、李珊珊、许允飞、韩军、陶一寒、米琳莹、王有芬、杨涵溪、杨丝丝对我的帮助，为我分摊许多运行工作的压力。感谢樊东卫、何勃亮在 BASS、DESI 数据上给予的支持。

感谢一起学习与研究的张静怡博士、康子涵、张震、张琦乾、吴莹、杨嘉宁、左肖雄、邵务俊、马鹏辉、朱嘉莹等提供的帮助。

感谢在国家天文台的学习与工作中经予我无私的指导和帮助的所有老师、同学和朋友。

感谢我的父母、岳父母、妻子多年来的默默支持，为我承担繁重的家务，始终不渝地给予我鼓励和信任。在我感到疲惫时，给予我家的温暖。他们是我前进道路上克服一切困难的力量之源。

感谢我孩子们的支持，正是由于你们的自律自强，我才能专心致志地从事科研工作。让我们一起朝着心中的梦想继续努力！

感谢我所有的亲人和朋友，感谢您们一直以来对我的关心、鼓励与帮助。

由于本人学识水平有限，在论文中难免有不足之处，恳请各位老师给予指正，深表感谢。

本论文完成之时，正值中国共产党第二十次全国代表大会召开。在这个并不安宁的世界，感谢祖国所给予的和平环境。

2022 年 12 月

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

作者简历：

学习经历：

2020 年 09 月——至今，在中国科学院国家天文台攻读博士学位。

2002 年 09 月——2005 年 06 月，在兰州大学信息学院获得硕士学位。

工作经历：

1995 年 07 月——1998 年 08 月，江西省遂川县堆前小学任教

2000 年 07 月——2002 年 08 月，江西省遂川县堆前中学任教

2005 年 08 月——2008 年 12 月，智尚网（北京）科技有限公司

2009 年 01 月——2011 年 10 月，北京晶合世纪数码科技有限公司

2011 年 11 月——至今，中国科学院国家天文台

已发表（或正式接受）的学术论文：

(1) **Changhua Li**, Yanxia Zhang, Chenzhou Cui, Dongwei Fan, Yongheng Zhao, Xuebing Wu, Boliang He, Yunfei Xu, Shanshan Li, Jun Han, Yihan Tao, Linying Mi, Hanxi Yang, Sisi Yang, Identification of BASS DR3 sources as stars, galaxies, and quasars by XGBoost, 2021, MNRAS, 506,1651-1664

(2) **Changhua Li**, Yanxia Zhang, Chenzhou Cui, Dongwei Fan, Yongheng Zhao, Xuebing Wu, Jing-yi Zhang, Jun Han, Yunfei Xu, Yihan Tao, Shanshan Li, Boliang He, Photometric redshift estimation of BASS DR3 quasars by machine learning, 2022, MNRAS, 509, 2289-2303

(3) **Changhua Li**, Yanxia Zhang, Chenzhou Cui, Dongwei Fan, Yongheng Zhao, Xuebing Wu, Jing-Yi Zhang, Yihan Tao, Jun Han, Yunfei Xu, Shanshan Li, Linying Mi, Boliang He, Zihan Kang, Youfen Wang, Hanxi Yang, Sisi Yang, Photometric Redshift Estimation of Galaxies in the DESI Legacy Imaging Surveys, 2023, MNRAS, 518, 513–525

(4) Yunfei Xu, Dong Xu, Chenzhou Cui, Dongwei Fan, Zipei Zhu, Bangyao Yu, **Changhua Li**, Jun Han, Linying Mi, Shanshan Li, Boliang He, Yihan Tao, Hanxi Yang, Sisi Yang, GWOPS：A VO-technology Driven Tool to Search for the Electromagnetic Counterpart of Gravitational Wave Event, 2020, PASP, 132, 104501

(5) 韩军、樊东卫、陶一寒、许允飞、李珊珊、米琳莹、**李长华**、崔辰州，FAST 科学观测项目管理信息系统，2022，数据与计算发展前沿，4，20-29

参加的研究项目及获奖情况：

1. 海量多波段天文数据融合关键技术与科学应用，科技部重点研发计划。