使用迁移学习方法寻找银道面背景类星体

傅煜铭（fuym@pku.edu.cn）

北京大学物理学院天文学系 / 北京大学科维理天文与天体物理研究所， 北京 100871

摘要：银道面背景类星体是重要的天体测量位置参考，以及研究银河系气体的探针。由于银道面聚集了银河系内的大部分恒星，并且存在比其他天区更为严重的尘埃消光，寻找银道面背景类星体一直十分困难。在银道面以外（高银纬）天区，机器学习方法广泛应用于类星体的选源研究中。而对于银道面天区的类星体选源，已有的机器学习模型不再适用，因为高银纬天区和银道面天区的天体测光数据遵循不同的统计分布。为了应对这种机器学习中的数据集偏移问题，本研究构建了一个迁移学习框架，从而通过测光数据寻找出银道面背景类星体：在数据层面，我们基于斯隆数字化巡天的星表和银河系消光图模拟了银道面背景的类星体和星系，从而合成出可用的训练集；在算法层面，我们将恒星—星系—类星体的三分类任务拆分为两次二元分类，从而降低类别不平衡与类别比例变化对于分类效果的影响。我们对具有Pan-STARRS1和ALLWISE测光数据的银道面天体（银纬范围：|b|≤20°）应用了XGBoost分类算法，并通过Gaia天体测量数据排除了大部分恒星污染。最终，本研究构建出包含157232个源的银道面背景类星体候选体表，样本测光红移范围在0到5之间。本研究将类星体的系统搜寻拓展到了密集星场，并且展示了使用天文领域知识提升数据挖掘效果的可行性。

一、项目背景

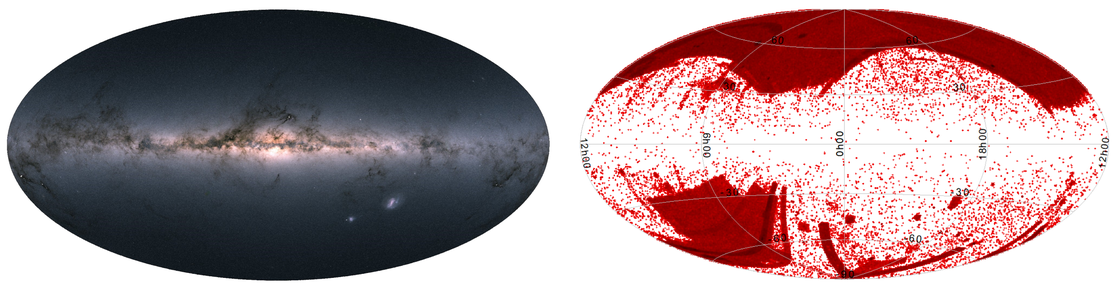

图1. 左：欧洲空间局（ESA）Gaia卫星拍摄的全天十七亿个天体，其中大部分为银河系内的恒星；右：截至2015年全天已发现的五十余万个类星体分布（Flesch, 2015）。两幅图均使用银道坐标系绘制，图像中部为银道面（Galactic plane）区域。

1. 重要的银道面背景类星体（Quasars behind the Galactic Plane; GPQs）：

- 重要的天体测量位置参考源（距离遥远，位置近乎不动）；
- 有效研究银河系气体的探针（利用光谱中的银河系吸收线）。

2. 寻找银道面背景类星体如同大海捞针：

- 银道面恒星十分密集，对背景天体测光数据造成污染；
- 银道面尘埃消光严重，造成背景天体显著变暗、变红；
- 银道面天区的天体和高银纬天体视星等、颜色分布不同，不同类别天体的比例也大不相同，“**数据集偏移**”不可忽视；
- 缺少银道面天区的类星体和星系训练样本。

二、解决方案：基于特征的迁移学习

1. 使用高银纬类星体、星系模拟对应的银道面背景天体，获得新的颜色-星等特征分布：

- 改正输入样本（SDSS类星体和星系）的银河系消光；
- 为输入样本随机产生银道面上的新坐标（|b|<20°）；
- 根据Planck卫星的消光图添加新坐标上的银河系消光；
- 选择亮于巡天极限星等的模拟样本作为优质模拟样本；
- 优质模拟样本+未选中的输入样本=机器学习训练样本。

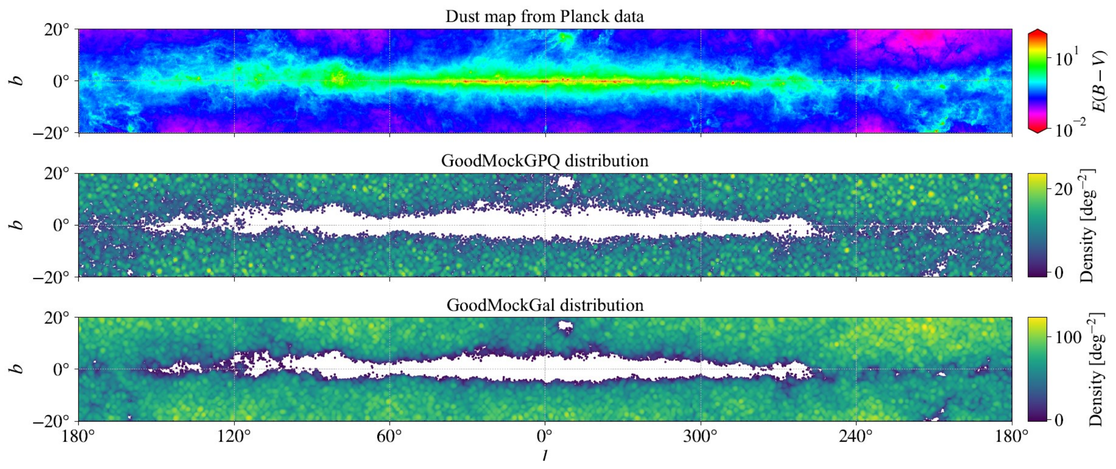

图2. 上：Planck卫星测量的银河系消光图（银道面|b|<20°部分）；中：模拟的银道面背景类星体密度分布；下：模拟的银道面背景星系密度分布。由于银道面中央消光过于严重，现有测光巡天无法探测到最中央的背景类星体和星系（中下两图的空白区域）。

2. 两步XGBoost机器学习二元分类，减轻类别不平衡影响：

- LAMOST银道面恒星加入训练样本；
- 第一步将天体分为恒星与河外天体（星系、类星体两个少数类合为一类），第二步将河外天体分类为星系与类星体，两次分类分别进行参数调优；
- 使用特征（颜色+形态，光学Pan-STARRS1+红外ALLWISE）：g-r, r-i, i-z, z-y, g-W1, r-W1, i-W1, z-W1, y-W1, W1-W2, W2-W3, i-i_kron, z-z_kron.

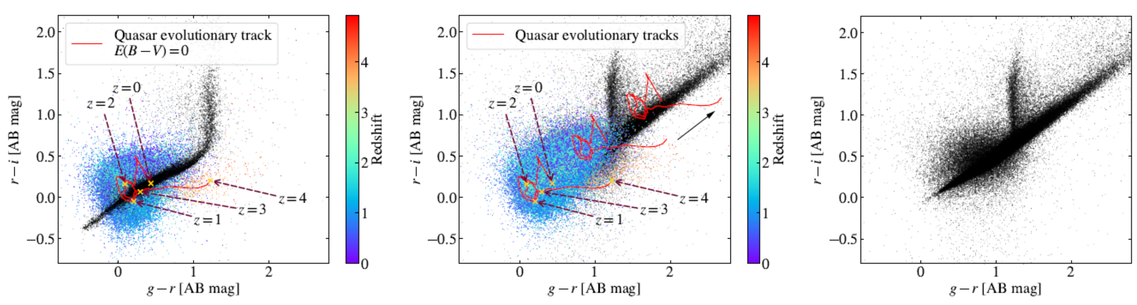

图3. r-i与g-r颜色—颜色图。左：高银纬的类星体（彩色）与恒星（黑色）；中：模拟的银道面类星体（彩色）与银道面点源（黑色，待分类，大部分为恒星）；右：同中间的银道面点源。类星体的颜色对应红移范围。左图展示了消光造成的红化值E(B-V)=0时的类星体颜色随红移演化轨迹，中图展示了E(B-V)分别等于0，0.75，1.5时的类星体颜色轨迹。随着消光（红化）增大，类星体颜色分布更加弥散，与恒星重叠更严重。

三、降低恒星污染，计算测光红移

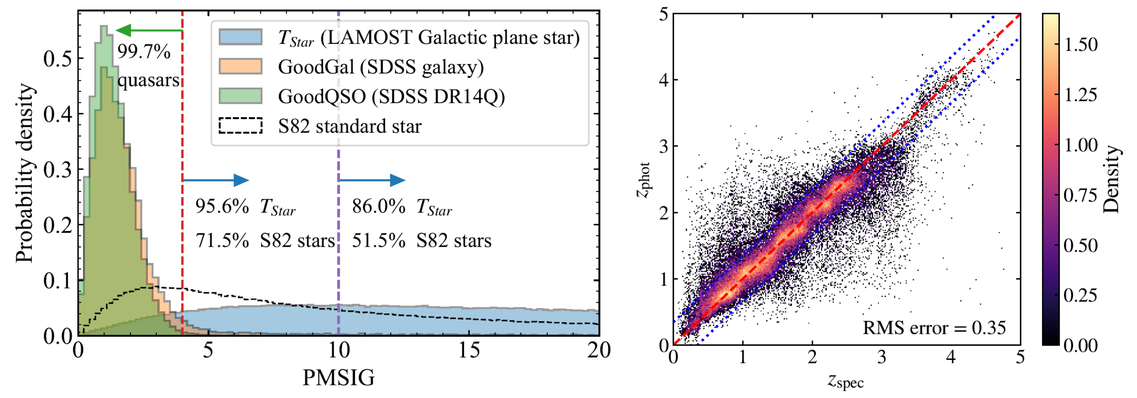

图4. 左：使用Gaia自行数据计算的自行显著程度PMSIG<=4可以有效选择大部分类星体，并去除大部分恒星；右：使用XGBoost回归算法基于SDSS类星体样本可以由测光数据估算出类星体的红移，验证集上的均方根误差约为0.35。

四、首个GPQ候选体表，近千源已获证认

1. 首个银道面背景类星体候选体表，共十五万余个源，恒星污染比例小于10%，红移范围宽（0<z<5）.

2. 近千个候选体通过光学望远镜（LAMOST等）光谱证认。

3. 后续：银道面天体测量系统误差分析、GPQ气体吸收线研究。

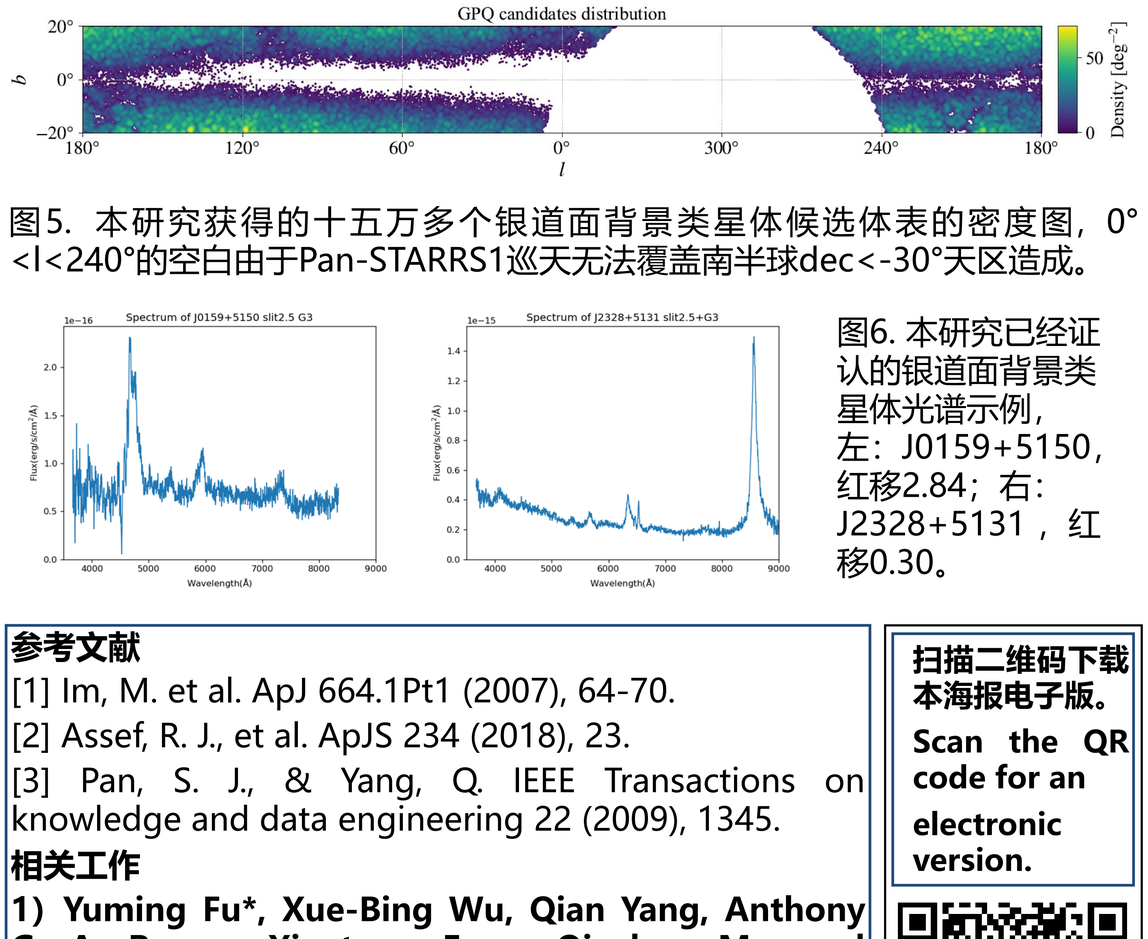

图5. 本研究获得的十五万多个银道面背景类星体候选体表的密度图，0°<l<240°的空白由于Pan-STARRS1巡天无法覆盖南半球dec<-30°天区造成。

图6. 本研究已经证认的银道面背景类星体光谱示例，左：J0159+5150，红移2.84；右：J2328+5131，红移0.30。

参考文献

[1] Im, M. et al. ApJ 664.1Pt1 (2007), 64-70.

[2] Assef, R. J., et al. ApJS 234 (2018), 23.

[3] Pan, S. J., & Yang, Q. IEEE Transactions on knowledge and data engineering 22 (2009), 1345.

相关工作

1) Yuming Fu*, Xue-Bing Wu, Qian Yang, Anthony G. A. Brown, Xiaotong Feng, Qinchun Ma, and Shuyan Li, ApJS.

IF:7.950 Submitted

*Email: fuym@pku.edu.cn

*Website: https://yumingfu.space/

扫描二维码下载本海报电子版。

Scan the QR code for an electronic version.