中国科学技术大学
University of Science and Technology of China

基于机器学习方法估计星系中性氢质量

答辩人：宋致远　　导师：孔旭教授

日期：2019年5月31日

目录

背景介绍

中性氢的重要性：

气体作为恒星形成的原料，是星系重要的物质组成部分。而中性氢是气体中重要组成部分，对中性氢成分的研究有助于我们理解星系的演化。

中性氢观测现状：

- 相比80、90年代有很大进步，有一批大天区的巡天观测，如HIPASS、ALFALFA。
- 相比光学等波段差距明显，观测数量少、观测深度浅。

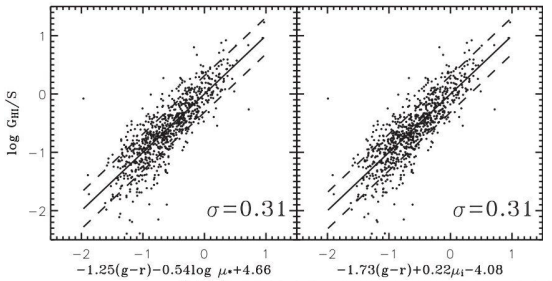
背景介绍

对中性氢质量与其他物理量关系的研究：

- **Zhang et al. 2009**

$$\log(M_{HI}/M_*) = -1.73(g-r) + 0.22\mu_i - 4.08$$

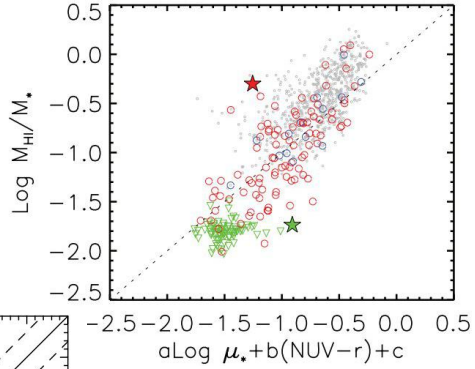
scatter=0.31dex



- **Catinella et al. 2010**

$$\log(M_{HI}/M_*) = -0.332\log\mu_* - 0.240(NUV-r) + 2.856$$

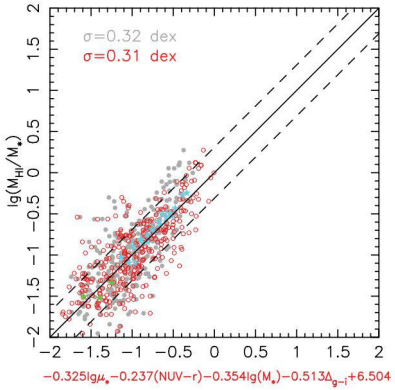
scatter=0.315dex



- **Li et al. 2012**

$$\log(M_{HI}/M_*) = -0.325\log\mu_* - 0.237(NUV-r) - 0.354\log M_* - 0.513\Delta_{g-i} + 6.504$$

scatter=0.31dex



线性模型，误差弥散始终大于0.3dex。

背景介绍

本文使用机器学习中的随机森林方法，利用非线性模型估计中性氢质量。

数据介绍

中国科学技术大学
University of Science and Technology of China

1.ALFALFA(Arecibo Legacy Fast ALFA Survey)

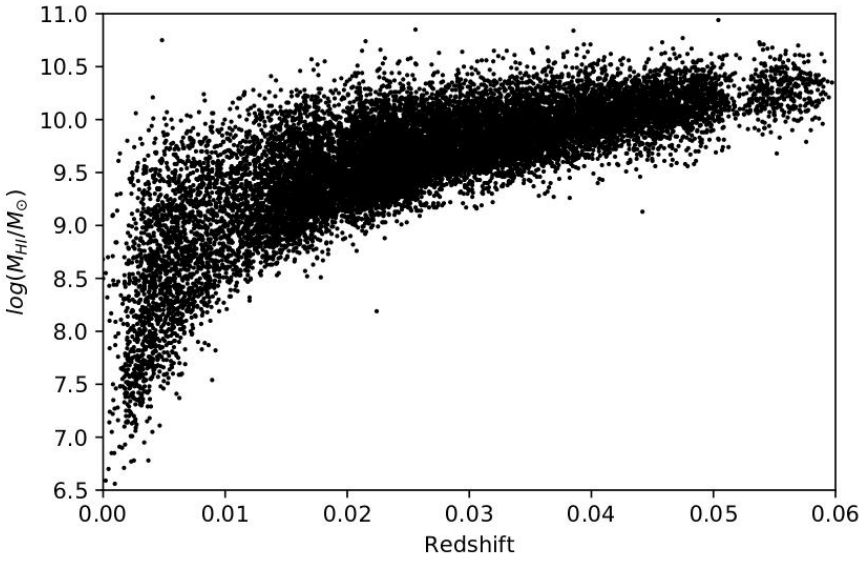
中性氢盲扫巡天，观测了红移0.06以内约31500个河外中性氢源

2.NSA(NASA-Sloan Atlas)

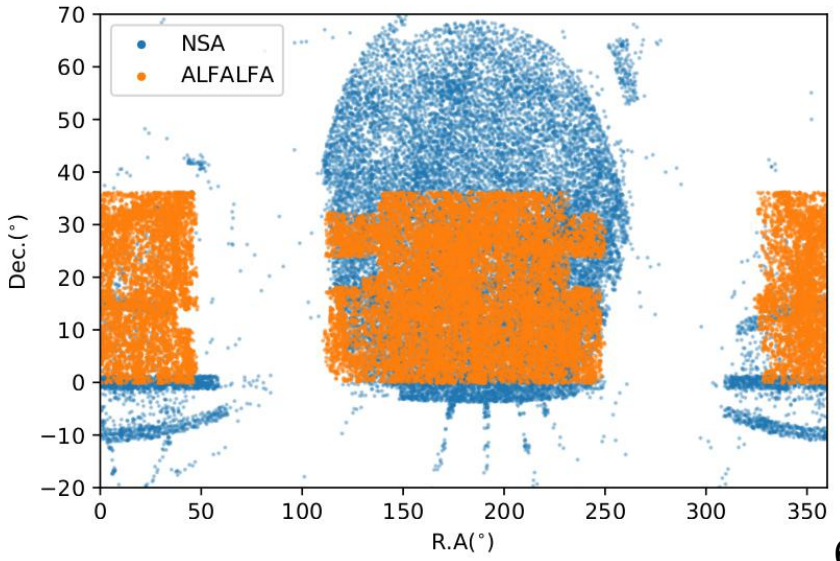
SDSS DR11+GALEX　　光学、紫外测光数据+恒星质量

3.GASS(GALEX Arecibo SDSS Survey)

0.025<z<0.05，1000个大质量星系的中性氢成分



ALFALFA中性氢质量-红移分布



ALFALFA、NSA天区分布

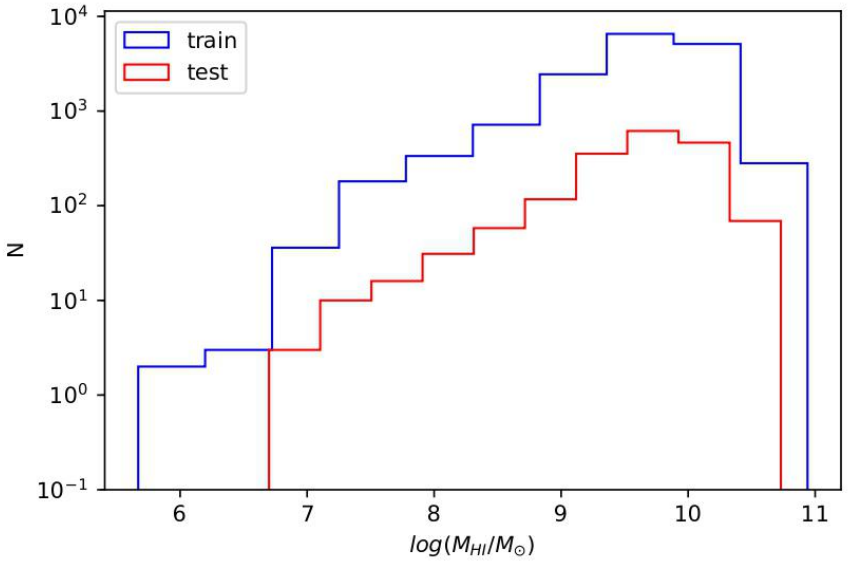
数据介绍

用于估计的物理量

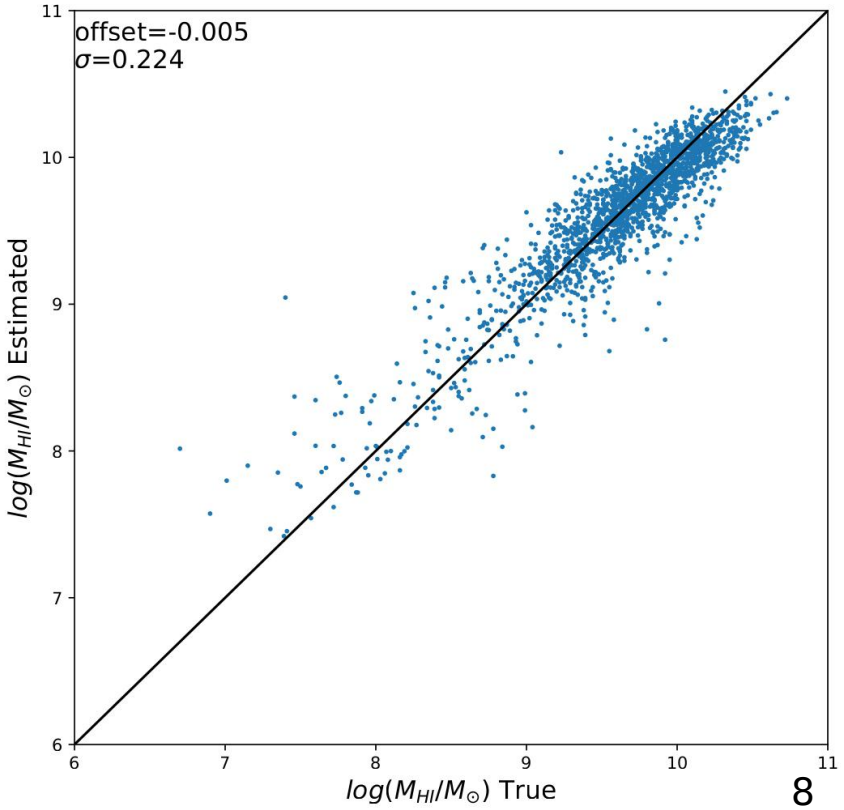
- 恒星质量
- R_{25}等照度半径
- 颜色（g-r,NUV-r)
- 聚集度
- Sersic指数
- i波段半光半径
- 质量面密度
- i波段面亮度
- 红移（仅用于分类)

预测模型

- 匹配ALFALFA和NSA
- 按照中性氢质量分层随机划分训练集和测试集
- 训练随机森林模型并用测试集检验



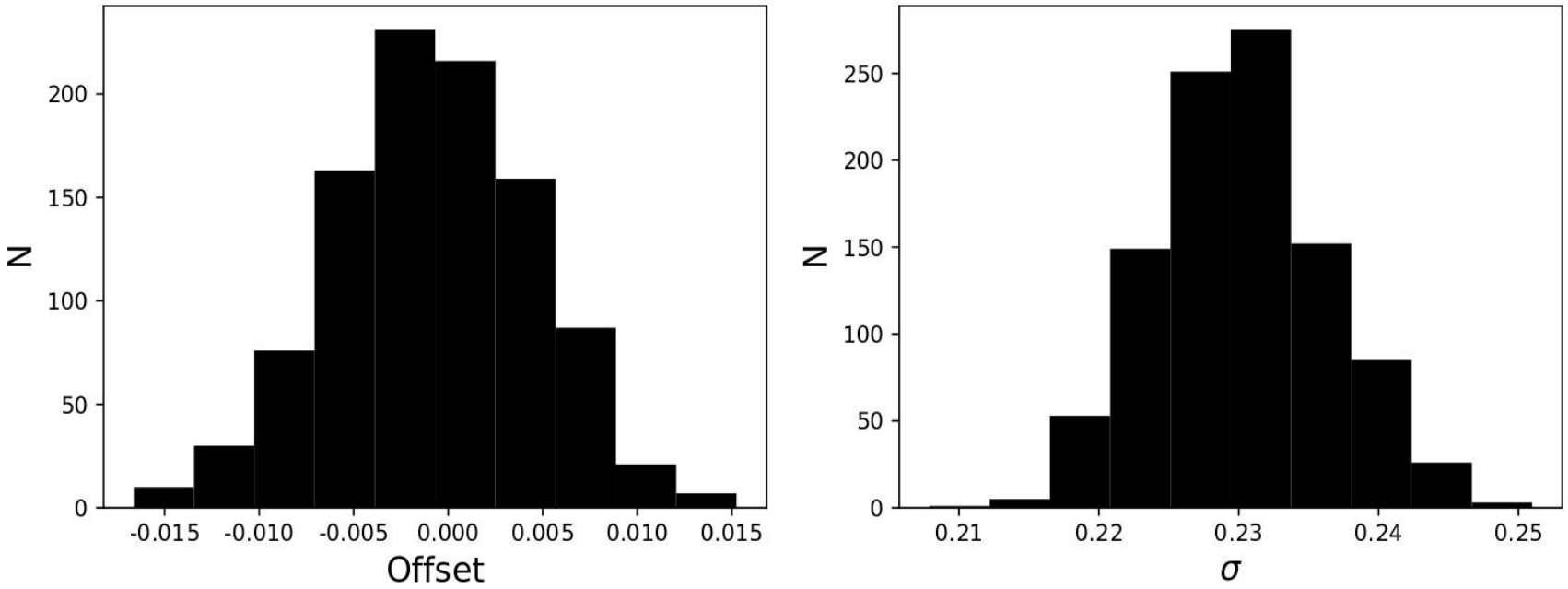
训练集、测试集中性氢质量分布



测试结果

预测模型

为避免单次划分的随机性，重复随机划分、训练、测试1000次。

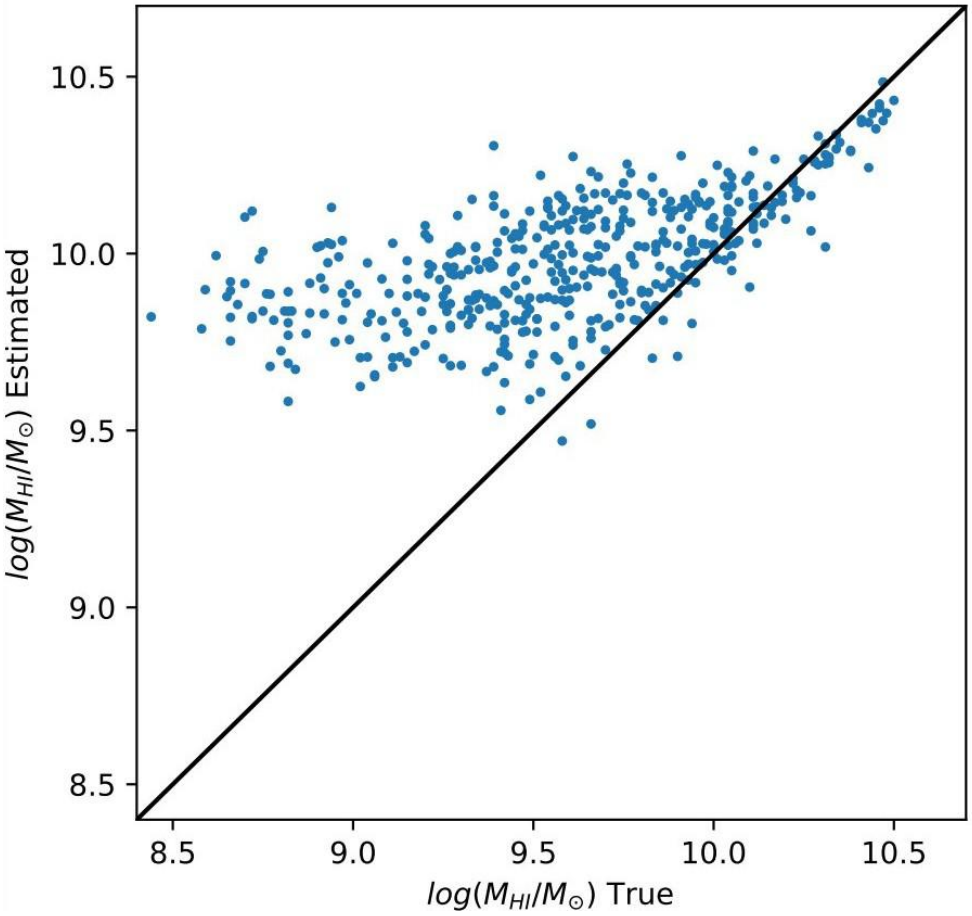


误差偏移、弥散分布

最终预测模型由全部匹配样本构成，其误差弥散应在0.23dex左右。

预测模型

预测模型应用于GASS，存在系统性偏差。



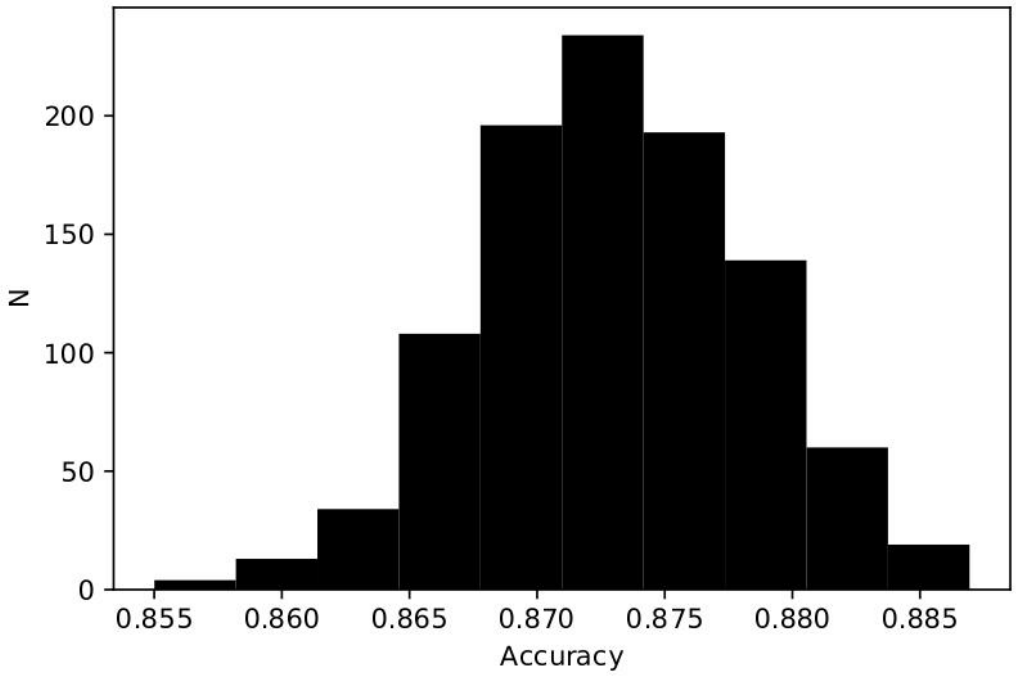
原因：GASS观测更深，用ALFALFA学得的预测模型会高估GASS样本的中性氢质量。

解决方法：判断星系是否适用预测模型。

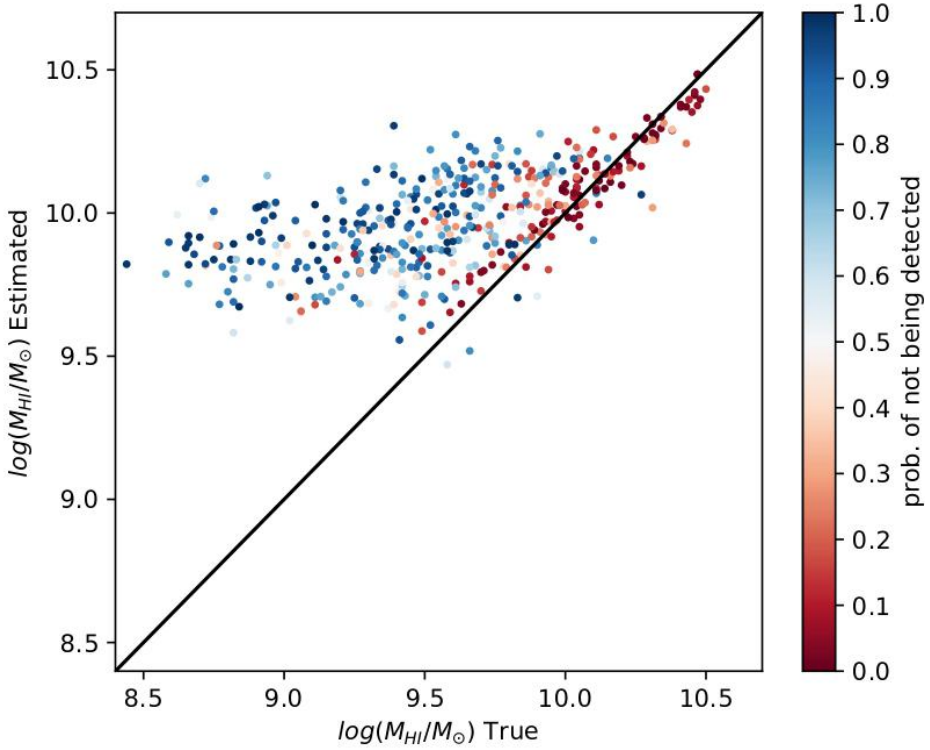
分类模型

适用的星系：ALFALFA可观测的星系
不适用的星系：ALFALFA不可观测的星系

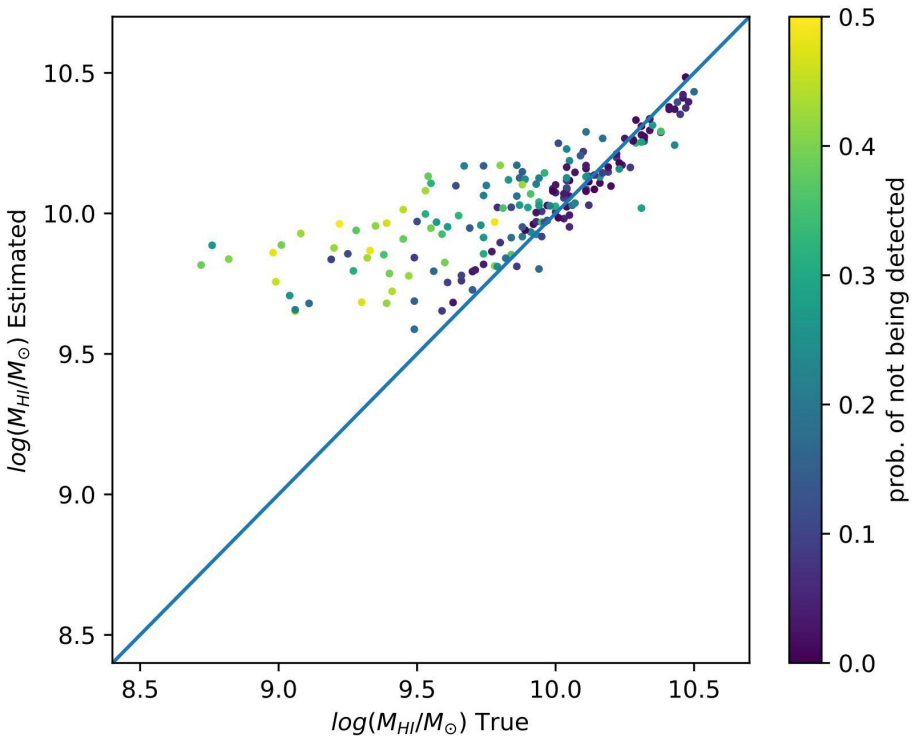
依然重复1000次划分、训练、测试，准确率分布如下：



分类模型



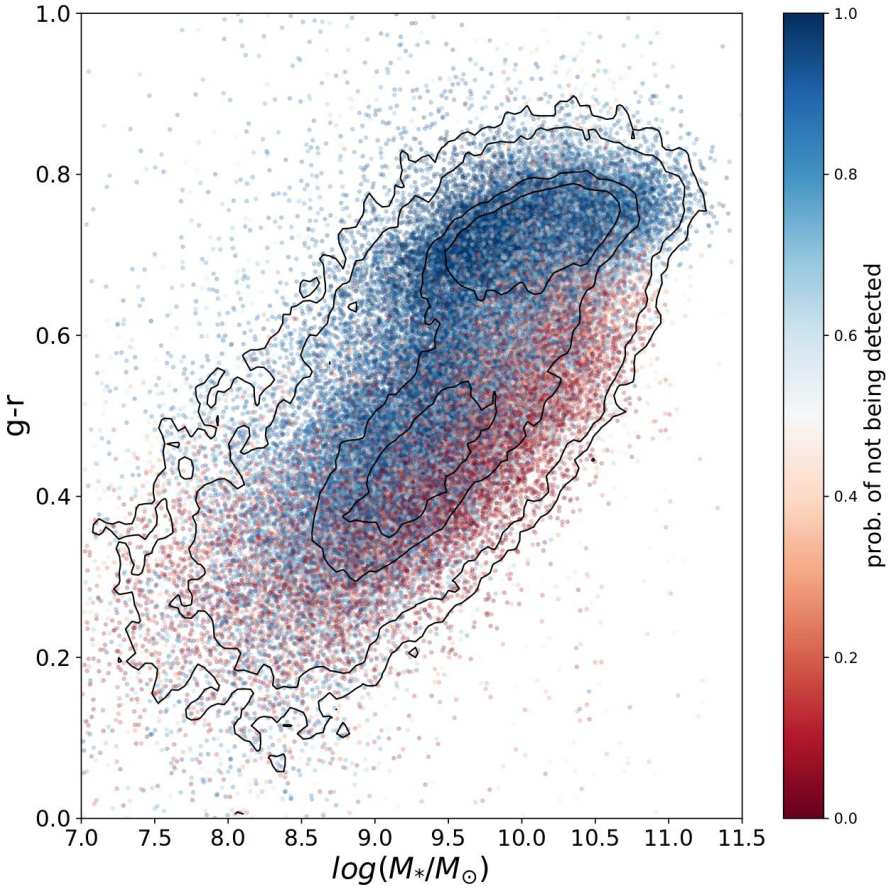
颜色代表随机森林的投票结果。偏差越大的星系，分类模型越不支持其适用预测模型。



应用分类模型后，大部分高估的星系被剔除，剩余少量高估的星系投票结果在50%左右，可能由于误判造成。

模型应用

对不在ALFALFA天区内的NSA星系使用分类模型

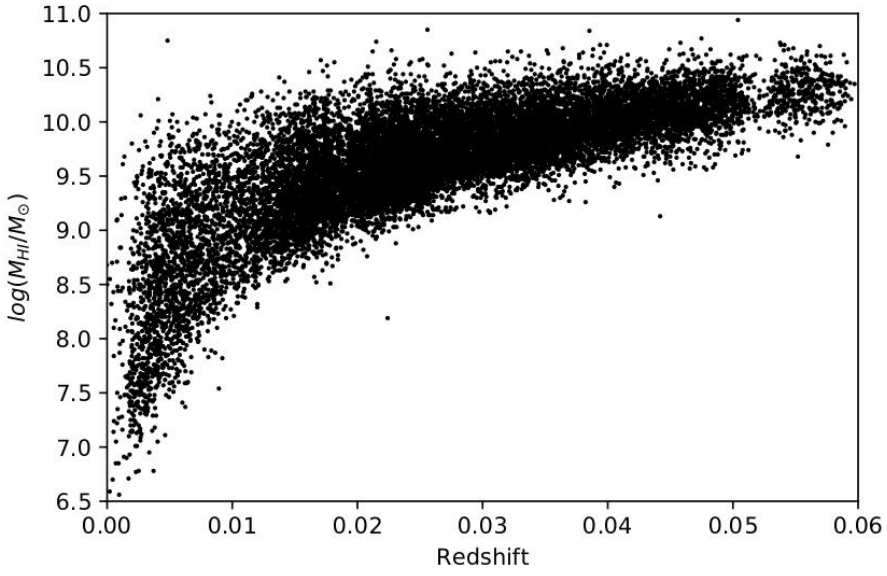


颜色红的星系大都无法预测；颜色蓝的星系大部分可以预测，一些小质量的星系无法预测。

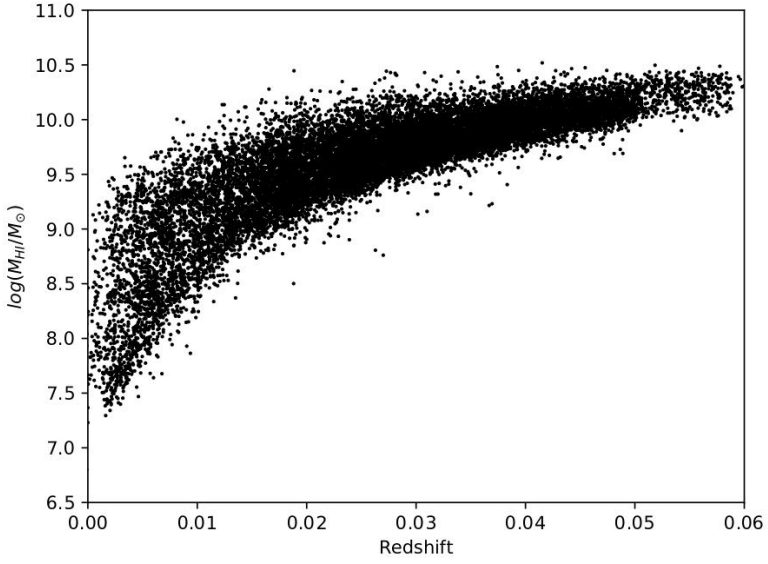
原因：红星系缺少气体，恒星形成活动弱；蓝星系气体较多，恒星形成剧烈，但小质量的星系中性氢质量也较小，在较高红移时可能无法观测。

模型应用

预测模型和分类模型结合，学习了ALFALFA盲扫的流量限特征，相当于对未观测的天区进行了一次观测。



ALFALFA样本中性氢质量-红移分布



可预测样本的中性氢质量-红移分布

总结展望

总结：

- 我们通过训练随机森林模型，降低了对中性氢质量估计的误差弥散。
- 通过分类模型，可以判断哪些星系可以适用我们的预测模型。从而对NSA中未观测天区的中性氢质量进行了估计。

展望：

- 两个模型的结合，本质上是对未观测的天区进行一次观测，如果期待提升预测能力，需要更多的观测数据支持。

谢谢！

附录

决策树

一种树形结构，其中每个内部节点表示一个属性上的判断，每个分支代表一个判断结果的输出，最后每个叶节点代表一种分类结果。

随机森林

集成模型，多个决策树结果的综合。

参考文献

Catinella B, Schiminovich D, Kauffmann G, et al. Apr 2010. The GALEX Arecibo SDSS Survey - I. Gas fraction scaling relations of massive galaxies and first data release[J]. MNRAS. 403(2):683–708.

Li C, Kauffmann G, Fu J, et al. Aug 2012. The clustering of galaxies as a function of their photometrically estimated atomic gas content[J]. MNRAS. 424(2):1471–1482.

Zhang W, Li C, Kauffmann G, et al. Aug 2009. Estimating the HI gas fractions of galaxies in the local Universe[J]. MNRAS. 397(3):1243–1253.