Semester Project Report (End-Sem) 2022

Classification of HI absorption spectra using ML

Nemmani Anirudh Srivastha

20191108
Indian Institute of Science Education and Research Tirupati

Under the supervision of Dr Arunima Banerjee

December 2, 2022

Abstract

HI - 21cm line in absorption spectra of continuum radio source can be observed when a gas medium associated or intervening in the line of sight of the source galaxy and the observer is present. The classification is usually done by looking at the optical spectrum of the source. Instead, we used Machine Learning to classify the spectra using line properties obtained from Busy Function fitting. We used Random Forest, KNN, Decision Tree, SVM and Logistic Regression models to train the dataset. We also used our dataset to try to train Neural Network and Convolution Neural Networks and discussed their difficulties. In this work, We obtained the highest accuracy and precision in classifying the spectra using the Random Forest model.

Keywords
HI Absorption spectra, Busy Function, Machine Learning

1 Introduction

Neutral Hydrogen is an important raw material for star formation. We probe its 21cm spectral line due to spin-flip transition to study HI. Radio waves can penetrate through dust clouds and reach the observer, which is why 21cm gives a better picture of Hydrogen in the galaxy or universe.

HI emission continuum can be used to understand the distribution of ISM in the galaxies[12]. We can study galaxies' evolution and kinematics, and it helps precisely measure cosmological constants[11].

It is noticed that the HI line in absorption due to a gas medium in the line of sight of the bright radio continuum is very effective in studying ISM properties at a much higher redshift compared to emission lines[10].

HI absorption spectra are classified into two groups, Associated and Intervening. If the absorbing gas medium is associated with the source galaxy, the absorption spectra due to the gas medium are called Associated spectra. Similarly, if the absorbing medium is intervening in the source's galaxy line of sight, the absorption spectra are called Intervening spectra. We can study ISM properties in our galaxy and other galaxies using Intervening absorption spectra, which are more reachable than emission spectra[8]. In the case of Associated spectra, we can observe the relation between the gas and the radio source as they are associated with the same galaxy[10].

It is challenging to classify the HI absorption spectrum without an optical spectrum into associated or intervening, as the spectrum itself has a redshift but cannot be distinguished if it is from associated or intervening gas. In some cases, an absorption spectrum was not classified until an optical spectrum had been obtained[1].

Due to fast rotating gas around AGNs, the associated absorption spectra have a broader width [7] compared to a narrow width profile of intervening spectra [6]. Using these properties, we can classify the spectra using machine learning. We derive the line properties of the spectrum using the busy function[13]. These line properties are used to train the ML models like Random Forest, KNN, SVM, Decision Tree and Logistic regression. Neural Networks and convolution neural networks are also trained in this project, but the model failed to learn due to the data sample, which will be discussed in later sections.

2 Data Collection

I received spectra from Dr Curran[2], Dr Rajeshwari Dutta[5, 4, 3] and Dr Maccagni[9]. After obtaining the spectra, I conducted a literature survey to find duplicates and ensure the data were in the same axis and units before fitting it. The data from Dr Maccagni lies in the redshift between 0.02 and 0.25, and data from Dr Curran lies in the redshift range $z > 0.01$.

After finding duplicates and removing the non-usable (Low Resolution) data, we obtained 99 Associated and 25 Intervening spectra. The data we obtained from Dr Curran is digitized data from articles. Due to this, some spectra were in low resolution, which resulted in no use for machine learning. So we excluded those spectra with no improvement in quality after another run of digitization.

As we can notice, there is a class imbalance in our dataset. Due to this reason, we have contacted Dr Kanekar for more data on intervening spectra and have yet to receive the data.

It should also be noted that the redshift distribution of known associated 21cm absorbers is $z < 1$, a majority in $z < 0.25$ [9]. At the same time, the intervening spectra have $z > 1$(conversation with Dr Kanekar). To make a statistical comparison, we assume no redshift evolution, which may not be true as the line properties such as velocity width and optical depth are likely to evolve intrinsically with redshift. At this point, we can only assume no redshift evolution and wait for more HI 21cm absorbers to be detected through blind surveys.

3 Analysis

3.1 Busy Function Fitting

In this project we use the busy function to fit the HI absorption spectra[13]. The busy function is given by the expression,

$$\begin{aligned} B_1(x) = \frac{a}{4} &\times (\mathrm{erf}[b_1\{w + x - x_e\}] + 1) \\ &\times (\mathrm{erf}[b_2\{w - x + x_e\}] + 1) \\ &\times (c|x - x_p|^n + 1) \end{aligned} \tag{1}$$

It has a maximum of 8 free parameters. To fit the spectra, we use BusyFit, written by Dr Tobias Westmeier[13]. One of the disadvantages of the busy function is that it only outputs positive values, evident from Eq.(1). After discussing with Dr Westmeier, it is noted that we can reverse the spectra for fitting it using the busy function under the assumption that it has a baseline centred on zero, which is the case for absorption spectra.

After using the busy fitting, we could fit a total of 103 spectra, losing fitting for 21 spectra due to poor resolution. Out of 21, 3 were from Intervening and 18 are from Associated spectra.

The busy function fitting parameters for each spectra were saved in my GitHub repository in the directory named "Final", the link for the repository is provided in the footnotes.[1]

3.2 Machine Learning Algorithm

We have used Random Forest, Decision Tree, KNN, SVM, Logistic Regression, Neural Network and Convolution Neural Network to classify the spectra using the line features obtained from fitting the spectra.

The features selected for the classification are the eight free parameters of busy function, along with centroid, w_{50}, w_{20}, peak flux, and integrated flux. We chose 13 features to classify the spectra.

The dataset is divided into ten subsets by randomly mixing the entire data and splitting them into training and test data. (i.e. ten-fold cross-validation). For necessary models, I have normalized the data inside the folds, so there will not be any information leakage to the test data set.

As we noticed in the previous section, there was a significant class imbalance in the dataset. To improve this major class imbalance, I have used SMOTE to over-sample the minority class so that the class imbalance will not affect the model.

After prepping the data, I used this data to train the data and tuning hyper-parameters for the machine learning model to improve the accuracy. The results section will discuss the obtained accuracy of each machine learning.

In the case of Neural Networks, it has been noticed that the model I have created was not learning, resulting in very low test accuracy. Even using SMOTE on minority class did not help the neural network learn as the dataset was small enough to train a good model. The same scenario is noticed in the case of a Convolution Neural Network.

Due to the stochastic nature of labeling the classes in GMM algorithm, GMM algorithm is not used in this project.

Please refer to my GitHub repository in the footnotes to access my machine-learning codes.[1]

4 Results

I have achieved the following Average Accuracy, AUC ROC and Average Precision for these machine learning models. I have used Zero Rule Model as a baseline to check my other machine learning algorithms, Zero Rule model predicts every sample into a single class, in this case it's the most frequent class.

- **Zero Rule Model** - Baseline Model (Predicts the same class (Most frequent class))
  - ROC AUC - 0.5
  - Average Accuracy - 78.7%
  - Average Precision - 21.3%
- **Random Forest**
  - ROC AUC - 0.969
  - Average Accuracy - 92.6%
  - Average Precision - 89.4%
- **KNN**
  - ROC AUC - 0.919
  - Average Accuracy - 86.3%
  - Average Precision - 79%
- **Decision Tree**
  - ROC AUC - 0.912
  - Average Accuracy - 93.2%
  - Average Precision - 79.3%

[1] The link to my GitHub Repository is Click Here. Note that this repository contains all my codes and bash scripts I have written to do all my analysis, and I will frequently push my edits throughout the project.

- **Logistic Regression**
  - ROC AUC - 0.916
  - Average Accuracy - 85.4%
  - Average Precision - 80%
- **SVM**
  - ROC AUC - 0.9
  - Average Accuracy - 83.8%
  - Average Precision - 75.3%

The above results may vary a little ($<$ 1%) due to the stochastic nature of oversampling.

5 Conclusion and Future Plans

We were able to train five different machine learning models and achieved the highest accuracy, precision and ROC AUC for Random Forest. All the machine learning models produced an average accuracy of above 80%. This shows that machine learning can classify the spectra as associated or intervening. These accuracies and precision can be improved with more data, reducing class imbalance and giving more features for the machine learning model. As mentioned in the Data Collection section, we have contacted Dr Kanekar to obtain more data for this project, which will reduce the class imbalance ratio and may help us achieve a higher accuracy level due to higher spectra resolution. We also aim to improve our Neural Networks by collecting more data in future.

6 Acknowledgements

I thank Dr Arunima and Dr Rajeshwari for clearing my doubts and providing me with data and guidance for this project. I also like to acknowledge Dr Maccagni and Dr Curran for providing me with data and Dr Tobias Westmeier for providing the BusyFit software and helping me with technical issues regarding the software to do this project effectively, and Adarsh Mahor from NIT Surat for his suggestions in improving Neural Network Model.

References

[1] J. R. Allison and E. M. et. al. Sadler. Discovery of hi gas in a young radio galaxy at z = 0.44 using the australian square kilometre array pathfinder. *Monthly Notices of the Royal Astronomical Society*, 453(2):1249–1267, 08 2015.

[2] S. J. Curran, S. W. Duchesne, A. Divoli, and J. R. Allison. A comparative study of intervening and associated H i 21-cm absorption profiles in redshifted galaxies. *Monthly Notices of the Royal Astronomical Society*, 462(4):4197–4207, 08 2016.

[3] R. Dutta, R. Srianand, N. Gupta, and R. Joshi. Hi 21cm absorption from z $\sim$ 0.35 strong Mg ii absorbers. *Monthly Notices of the Royal Astronomical Society*, 468(1):1029–1037, mar 2017.

[4] R. Dutta, R. Srianand, N. Gupta, R. Joshi, P. Petitjean, P. Noterdaeme, J. Ge, and J.-K. Krogager. Incidence of hi 21-cm absorption in strong fe ii systems at 0.5 $<$ z $<$ 1.5. *Monthly Notices of the Royal Astronomical Society*, 465(4):4249–4264, nov 2016.

[5] R. Dutta, R. Srianand, N. Gupta, E. Momjian, P. Noterdaeme, P. Petitjean, and H. Rahmani. H i 21-cm absorption survey of quasar-galaxy pairs: distribution of cold gas around z $<$ 0.4 galaxies. *Monthly Notices of the Royal Astronomical Society*, 465(1):588–618, 10 2016.

[6] N. Gupta, R. Srianand, P. Petitjean, P. Noterdaeme, and D. J. Saikia. A complete sample of 21-cm absorbers at z $\sim$ 1.3: Giant metrewave radio telescope survey using mg ii systems. *Monthly Notices of the Royal Astronomical Society*, 398(1):201–220, 08 2009.

[7] J. Holt, C. N. Tadhunter, and R. Morganti. Fast outflows in compact radio sources: evidence for AGN-induced feedback in the early stages of radio source evolution. *Monthly Notices of the Royal Astronomical Society*, 387(2):639–659, 05 2008.

[8] N. Kanekar and F.H. Briggs. 21-cm absorption studies with the square kilometer array. *New Astronomy Reviews*, 48(11):1259–1270, 2004. Science with the Square Kilometre Array.

[9] Maccagni, F. M., Morganti, R., Oosterloo, T. A., Geréb, K., and Maddox, N. Kinematics and physical conditions of hn nearby radio sources - the last survey of the old westerbork synthesis radio telescope. *A&A*, 604:A43, 2017.

[10] Raffaella Morganti and Tom Oosterloo. The interstellar and circumnuclear medium of active nuclei traced by h i 21 cm absorption. *The Astronomy and Astrophysics Review*, 26(1), jul 2018.

[11] S. Rawlings, F.B. Abdalla, S.L. Bridle, C.A. Blake, C.M. Baugh, L.J. Greenhill, and J.M. van der Hulst. Galaxy evolution, cosmology and dark energy with the square kilometer array. *New Astronomy Reviews*, 48(11):1013–1027, 2004. Science with the Square Kilometre Array.

[12] G W Rougoor and J H Oort. Distribution and motion of interstellar hydrogen in the galactic system with particular reference to the region within 3 kiloparsecs of the center. *Proc. Natl. Acad. Sci. U. S. A.*, 46(1):1–13, January 1960.

[13] T. Westmeier, R. Jurek, D. Obreschkow, B. S. Koribalski, and L. Staveley-Smith. The busy function: a new analytic function for describing the integrated 21-cm spectral profile of galaxies. *Monthly Notices of the Royal Astronomical Society*, 438(2):1176–1190, dec 2013.