RNAAS RESEARCH NOTES OF THE AAS

The Application of Machine Learning to Quasar and Seyfert Classification

To cite this article: Vivek Abraham *et al* 2024 *Res. Notes AAS* **8** 46

Draft version February 2, 2024
Typeset using LATEX default style in AASTeX631

The Application of Machine Learning to Quasar and Seyfert Classification

Vivek Abraham,[1] Joel Deville,[1] and Garv Kinariwala[1]

[1] *University of Texas at Austin*

ABSTRACT

Machine learning can be utilized to classify spectra flagged as Active Galactic Nuclei (AGN) belonging to Seyferts or Quasars, expediting data collection and aiding in analyzing the AGN types. While many properties of Seyferts and Quasars can be used as feature points in training a machine learning model, one relatively available property with high information density is the spectra of the AGN types. This paper aims to describe the training and results of a K-Nearest Neighbors (KNN) and a Dense Neural Network (DNN) machine learning model built to classify AGNs as Seyfert type 1s, Seyfert type 2s, or Quasars.

1. INTRODUCTION

Since their discovery, two types of Active Galactic Nuclei (AGN), Seyferts and Quasars, have gained significant traction in the scientific community. Seyferts appear as very luminous celestial objects, with their total radiation output rivaling the amount emitted by the entirety of the constituent stars in their host galaxies. Seyfert objects can be further classified into two major classes based on the relative line widths of their emission spectra: Seyfert 1s and Seyfert 2s. While both Seyfert types have broad line widths relative to non-AGN spectra, Seyfert 1s have very broad permitted lines on the order of 10^4 kilometers per second, which Seyfert 2s lack. Like Seyferts, Quasars emit tremendous amounts of radiation; however, while Seyferts emit radiation comparable in intensity to their host galaxy, quasars outshine them immensely as they put off more light at further distances generally. Both Seyferts and Quasars have been attributed to be hosts of supermassive black holes in the center of galaxies, making them hot topics of scientific research. Understanding the processes behind Seyfert and Quasars will help expand our understanding of galaxy morphology, star formation, and the composition of the early universe. It is useful to consider other studies, which used similar machine learning methods, such as Ma et al. (2019) and Cavuoti et al. (2013). By looking at these studies and our own, we believe machine learning is crucial to accelerate the classification of AGNs based on the spectral values measured. We seek to fit a model that is able to accurately classify Seyferts and Quasars across varying redshifts, which proves difficult normally. This model will focus on doing spectral classifications of Quasars versus Seyferts.

2. DATA COLLECTION

Our study primarily utilized data from the Sloan Digital Sky Survey (SDSS) Data Release 18 (Kollmeier et al. 2019) and the Veron Catalog of Quasars & AGN (VeronCAT), 13th edition (Véron-Cetty & Véron 2010). We specifically retrieved data using SQL queries over data from SDSS-V and the Science Archive Server for FITS files. We were able to interface with the SDSS data through the Catalog Archive Server (CAS). CAS provided a vast array of objects with detailed spectra, notably from its SpecObj subset which includes about 5.1 million objects. However, as SDSS's classification of AGNs is limited, we incorporated the VeronCAT Catalog for more nuanced AGN classifications. We utilized TopCAT to cross-match a batch SQL query from SDSS with a TAP query from VeronCAT using a maximum

akka.viv@gmail.com

joeldeville114@gmail.com

garvkinariwala@gmail.com

error margin of 8 arc-seconds. After cross-matching, we had 61,270 data points, out of which approximately 35,000 spectra FITS files were available from the SDSS Science Archive Server.

We then processed this data with a Python program; This program extracted spectral information and downloaded spectra files. The Python program created a CSV file from this data collected, which we later used with Pandas to serve as the input to our model. The final analysis, involving trend identification and visualization, was performed using Python libraries such as Pandas and Matplotlib. This approach enabled us to efficiently categorize and analyze the data, particularly focusing on AGN subclassifications and their properties.

3. MODEL PARAMETERS

For the dense neural network, we used Tensorflow as the machine learning library. The training parameters were approximately 3900 measured flux values for different wavelengths, extracted from the FITS file, as well as the estimated redshift for the object. The training target was the category of the object (Quasar, Seyfert 1, or Seyfert 2). We used one-hot encoding for our three classification classes. This converted our categorical data of classifications into numerical values to be used as input and output for the dense neural network. For the hidden layers, we decided to use the ReLu activation function, which was found to be the best general activation function for dense neural networks in most cases (Bai 2022). The softmax activation function was utilized for the output layer for an easy conversion into labels. We decided to test two different hidden layer patterns: a 256-32 layer pattern and a 1024-256-64-16 layer pattern. For both patterns, we trained for 200 maximum epochs with early stopping enabled with a patience of 40. Finally, we tested our model both with L2 regularization implemented and without regularization, creating a total of four model architectures for the neural network; To reduce variation in our results, we tested each architecture three times with different random states and averaged the results together. For the KNN model, we tested four different nearest-neighbor sizes: 5 neighbors, 10 neighbors, 20 neighbors, and 50 neighbors. For each size, we tested a uniform and an inverse-to-distance weight function for prediction. Similar to the neural network test, we ran each subtest three times with different random states. To calculate accuracies, we used the Scikit-learn accuracy score metric.

KNN Results				
	5 Neighbors	10 Neighbors	20 Neighbors	50 Neighbors
Distance	91.39%	91.76%	91.41%	89.88%
Uniform	91.76%	91.40%	91.00%	90.12%

Neural Network Results		
	Small	Large
No Regularization	93.16%	93.98%
Regularization	92.18%	92.36%

Class Accuracy Breakdown			
	Quasars	Seyfert 1s	Seyfert 2s
10 Neighbor Distance KNN	98.89%	56.85%	79.06%
Non-regularized Large DNN	98.73%	71.97%	76.78%

Table 1. Comparison of Model Accuracies

4. RESULTS

The majority of the differences in predictions were from the KNN mispredicting true Seyfert 1s as Quasars, which the neural network was better at predicting. Another larger source of differences was with one model predicting an AGN as a Seyfert 1 and the other model predicting an AGN as a Seyfert 2. Both models generally agreed on Quasar predictions; moreover, an AGN classified as a Quasar by one model was seldom classified as a Seyfert 2 by the other model.

The kNN model tended to perform better with correctly predicting Seyfert 1s with lower redshifts relative to predicting Seyfert 1s with higher redshifts. This is likely due to Quasar objects tending to have a higher redshift value than Seyferts, allowing the model to use the redshift values to distinguish Quasars from Seyferts; moreover, high redshifts compress the spectra of objects, further aiding with differentiation. We hypothesize that the neural network performed better with predicting even high-redshift Seyfert 1s due to the neural network suffering less from the very high dimensionality of our training set (University 2018). The KNN model performed significantly worse than our dense neural network when dealing with Seyfert 1s. However, both the KNN model and the neural network performed worse on intermediate Seyfert subclassifications than their respective Quasar classifications; Subclassifications which gave our models issues were Seyfert 1.2s (which we grouped as a Seyfert 1), Seyfert 1.8s (which we grouped as a Seyfert 2), and especially Seyfert 1.5s (which we grouped as a Seyfert 1). Attempting to group Seyfert 1.5s as Seyfert 2s instead did not seem to improve the accuracy of our models, with the large, non-regularized neural network performing at 93.68% accuracy. A further breakdown on the best-performing models and information about other models can be found in Table 1.

We hypothesize that the greater degree of mispredictions for these Seyfert subclasses is due to the spectra of intermediate Seyferts being more ambiguous, containing features that are not distinctly similar to those of either true Seyfert 1s or Seyfert 2s. Our results corroborate with the unification scheme for Active Galactic Nuclei, which suggests that all AGN are formed through the same cosmic phenomenon (Spinoglio & Fernandez-Ontiveros 2019). The unification scheme theory implies that AGN features occupy a continuous range instead of distinct types, which means that some nuclei can't be easily classified by a machine learning model. It is worthwhile to note this theory is still a topic under debate, and the current inability to accurately classify intermediate classes of objects does lend support to the notion of AGN properties being on a spectrum.

We want to thank Dr. Mitra, Dr. Gebhardt, our peer mentors, and all members of the Geometry of Space stream in the Freshman Research Initiative at the University of Texas at Austin for all the help they gave along the way.

REFERENCES

Bai, Y. 2022, RELU-Function and Derived Function Review, ,

Cavuoti, S., Brescia, M., D'Abrusco, R., Longo, G., & Paolillo, M. 2013, Photometric Classification of Emission Line Galaxies with Machine-Learning Methods, ,

Kollmeier, J., Anderson, S. F., Blanc, G. A., et al. 2019, SDSS-V Pioneering Panoptic Spectroscopy, ,

Ma, Z., Xu, H., Zhu, J., et al. 2019, A Machine Learning Based Morphological Classification of 14,245 Radio AGNs Selected from the Best-Heckman Sample, ,

Spinoglio, L., & Fernandez-Ontiveros, J. A. 2019, AGN Types and Unification Model, ,

University, C. 2018, Lecture 2: k-nearest neighbors / Curse of Dimensionality, ,

Véron-Cetty, M. P., & Véron, P. 2010, A catalogue of quasars and active nuclei: 13th edition, , , doi:10.1051/0004-6361/201014188