RNAAS RESEARCH NOTES OF THE AAS

Classifying Emission-line Galaxies Using a Dense Neural Network and Support Vector Machine

To cite this article: Ayan Gupta *et al* 2024 *Res. Notes AAS* **8** 82

Draft version March 13, 2024
Typeset using LaTeX default style in AASTeX631

Classifying Emission-Line Galaxies using a Dense Neural Network & Support Vector Machine

Ayan Gupta,[1] Jazhiel Segura-Monroy,[1] and Yash Totlani[1]

[1] *The University of Texas at Austin*

ABSTRACT

In this study we present an innovative approach to classifying emission-line galaxies, specifically categorizing them as Star-Forming, Seyfert, LINERs (Low Ionization Nuclear Emission Line Region), or Composites. Leveraging both a Dense Neural Network (DNN) and Support Vector Machine (SVM), we use key emission-line flux ratios as input features extracted from the Baryon Oscillation Spectroscopic Survey (BOSS) data within the Sloan Digital Sky Survey (SDSS). The high accuracy in classification for both Machine Learning models showcases their effectiveness and viability in accurately classifying emission-line galaxies with slightly different inputs and target classifications compared to past Machine Learning models.

1. INTRODUCTION

Only in the 20th century were emission lines initially identified within galactic nuclei, giving rise to the research field of emission-line galaxies (Shields 1999). By studying emission spectra, we can determine the composition and predominant ionizing source to distinguish between different types of emission line galaxies. The Baldwin-Phillips-Terlevich diagram (BPT), initially introduced in 1981, served as the pioneering tool for this purpose and underwent continuous advancements and modifications over time (Baldwin et al. 1981). With the advent of extensive data sets and improvements in computation, our research aims to enhance and scale the categorization of emission line galaxies using popular Machine Learning techniques. Our aim is to create models that match the performance of previous models used for categorizing emission-line galaxies, such as the SVM presented in Shi et al. (2015) and the different classifiers in Zhang et al. (2019).

2. METHODS

Effectively classifying galaxies through supervised learning techniques requires two key components: precise measurements of emission-line fluxes and dependable galaxy classification labels. For this purpose, spectroscopic data for galaxies were taken from observations of the SDSS/Baryonic Oscillation Spectroscopic Survey collaboration from SDSS Data Release 18 (Almeida et al. 2023). About 1.5 million galaxies with redshift values approximately to 0.7 have been recorded as part of the survey (Thomas et al. 2013). From the emission-line flux measurements provided by the dataset, we were able to construct reliable flux ratio values of [NII]/[Hα], [OIII]λ5006/[Hβ], [OII]/[OIII]λ5006, [OI]/[Hα], [SII]/[Hα], [OIII]λ5006/[OIII]λ4363, [OII]/[Hβ], that served as the input features to our classification algorithms. The selection of these specific ratios was based on their established utility in the classification of emission-line galaxies, as demonstrated in previous research diagrams. The dataset included galaxy classifications, drawing from the those presented in works by Kauffmann et al. (2003), Kewley et al. (2001), and Schawinski et al. (2007). Notably, at lower redshifts these traditional diagnostic diagrams demonstrate heightened reliability in facilitating accurate galaxy classifications (Zhang et al. 2019). Additionally, the use of a low redshift range improves completeness in the LINER class, given most LINERs are found at lower redshifts compared to Seyfert galaxies (Kewley et al. 2006). Therefore, we chose to acquire three distinct data sets, each comprising spectroscopic data and associated labels for galaxies within red shift intervals below 0.10, 0.15, and 0.30. The data sets for intervals below 0.15 and 0.30 include objects from the lower redshift ranges, ensuring a comprehensive coverage across the specified spectrums. The purpose is to train the model on data below 0.10 redshift and evaluate its performance on higher redshifts. Both models were trained and tested on the same data, facilitating a thorough evaluation of their performance.

Using TensorFlow and Keras (Abadi et al. 2015; Chollet et al. 2015), we designed a Dense Neural Network (DNN) for our galaxy dataset, featuring an input layer, three dense layers with 16, 64, and 32 neurons, and ReLU activation.

L2 regularization (strengths: 0.001, 0.01, 0.02) was applied to prevent overfitting. The output layer used soft-max for multi-class classification. The Adam optimizer (learning rate: 0.001) and categorical cross-entropy were selected. Training ran for 10 epochs. Hyper-parameter tuning via Keras-tuner (O'Malley et al. 2019) involved a grid-search on neurons (16-128), layers (1-10), learning rates (0.001-0.1), epochs (10-50), and regularization rates (0.001-0.03).

The SVM implementation utilized the Support Vector Classification (SVC) module from the scikit-learn library (Pedregosa et al. 2011). This model was configured with a radial basis function (RBF) kernel, a regularization parameter (C) set to 10, and a gamma value of 1. Data preprocessing included feature normalization on a scale of -1 to 1, with Keras-tuner again being used for hyperparameter tuning.

3. RESULTS

At $z < 0.10$, the final sample consisted of 90,281 galaxies, which was split 80/20 into a training and validation sample (72,225 galaxies) and a test sample (18,056 galaxies). The sample consisted of 61,960 star-forming galaxies, 15,391 composite galaxies, 9,185 LINERs, and 3,745 Seyferts. At $z < 0.15$, the sample consisted of 133,980 galaxies with 82,376 star-forming galaxies, 26,703 composite galaxies, 17,155 LINERs, and 7,746 Seyferts. The $z < 0.30$ sample consisted of 173,218 galaxies with 95,452 star-forming galaxies, 37,185 composite galaxies, 27,941 LINERs, and 12,640 Seyferts.

Table 1. Precision values for each galaxy class and overall accuracy for DNN & SVM

Dense Neural Network				SVM			
z < 0.10				**z < 0.10**			
Star-Forming	Composite	LINER	Seyfert	Star-Forming	Composite	LINER	Seyfert
98%	93%	96%	95%	95%	91%	94%	99%
Accuracy: 95.99%				**Accuracy: 94.16%**			
z < 0.15				**z < 0.15**			
Star-Forming	Composite	LINER	Seyfert	Star-Forming	Composite	LINER	Seyfert
98%	94%	96%	96%	95%	88%	92%	99%
Accuracy: 96.57%				**Accuracy: 93.17%**			
z < 0.30				**z < 0.30**			
Star-Forming	Composite	LINER	Seyfert	Star-Forming	Composite	LINER	Seyfert
98%	94%	95%	96%	95%	87%	91%	99%
Accuracy: 96.24%				**Accuracy: 92.79%**			

Employing accuracy as the discriminating criterion between the two models, the DNN emerges as a superior classification algorithm, demonstrating consistently higher accuracy across all redshift ranges. Notably, when assessing the DNN performance at higher redshifts, the accuracy exhibits very similar results suggesting that the emission-line ratios used in the model are also valuable for classifying emission-line galaxies at greater redshifts. The SVM also proves to be a viable classifier, demonstrating strong accuracy and precision metrics and only slight decreases at greater redshifts.

In the context of the findings presented in Zhang et al. (2019), our investigation reveals that our Deep Neural Network (DNN) and Support Vector Machine (SVM) display a considerable increase in accuracy compared to the K-Nearest Neighbors (KNN), Neural Network, Support Vector Classifier (SVC), and Random Forest outlined in their study. It is important to note, however, that the comparative analysis must take into consideration the broader redshift range employed in their research, a factor that inherently complicates the classification of galaxies. Moreover, despite our model achieving a marginally lower accuracy compared to the one introduced in Shi et al. (2015), it distinguishes itself by classifying objects into more specific categories.

4. CONCLUSION

The utilization of these models emerges as a robust alternative to previous models designated for the same task. Expanding upon the current research, additional refinements may involve the subdivision of galaxies into more detailed classifications, such as distinguishing between Seyfert 1 and Seyfert 2 galaxies. Moreover, there is an opportunity to delve deeper into the investigation of feature importance by exploring the inclusion or exclusion of various emission-line flux ratios.

5. ACKNOWLEDGEMENTS

Acknowledgments go to The University of Texas at Austin, the Freshman Research Initiative (FRI) Program and UT-Austin faculty, Dr. Shyamal Mitra and Dr. Karl Gebhardt.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., et al. 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/

Almeida, A., Anderson, S. F., Argudo-Fernández, M., et al. 2023, doi: 10.48550/ARXIV.2301.07688

Baldwin, J. A., Phillips, M. M., & Terlevich, R. 1981, 93, 5, doi: 10.1086/130766

Chollet, F., et al. 2015, Keras, https://keras.io

Kauffmann, G., Heckman, T. M., Tremonti, C., et al. 2003, 346, 1055, doi: 10.1111/j.1365-2966.2003.07154.x

Kewley, L. J., Dopita, M. A., Sutherland, R. S., Heisler, C. A., & Trevena, J. 2001, 556, 121, doi: 10.1086/321545

Kewley, L. J., Groves, B., Kauffmann, G., & Heckman, T. 2006, 372, 961, doi: 10.1111/j.1365-2966.2006.10859.x

O'Malley, T., Bursztein, E., Long, J., et al. 2019, Keras Tuner, https://github.com/keras-team/keras-tuner

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Journal of Machine Learning Research, 12, 2825

Schawinski, K., Thomas, D., Sarzi, M., et al. 2007, 382, 1415, doi: 10.1111/j.1365-2966.2007.12487.x

Shi, F., Liu, Y.-Y., Sun, G.-L., et al. 2015, 453, 122, doi: 10.1093/mnras/stv1617

Shields, G. 1999, 111, 661, doi: 10.1086/316378

Thomas, D., Steele, O., Maraston, C., et al. 2013, 431, 1383, doi: 10.1093/mnras/stt261

Zhang, K., Schlegel, D. J., Andrews, B. H., et al. 2019, 883, 1163–1176, doi: 10.3847/1538-4357/ab397e