

A Comprehensive Survey of Emission-Line Galaxy Classification

AYAN GUPTA,¹ JAZHIEL SEGURA-MONROY,¹ AND YASH TOTLANI¹

¹*University of Texas at Austin
110 Inner Campus Drive
Austin, TX 78712, USA*

(Dated: September 6, 2023)

ABSTRACT

In this paper, we give a brief survey about our research study: Using machine learning models to cluster Emission-Line from Galaxies. A overview of the problem at hand is given and its relevance in the larger context of astronomical research. Continuing, we give an overview of the current research done in this field so far as well as discuss what emission line galaxies are along with data collection techniques related to it. BPT Diagrams and other traditional models are discussed as well as modern day machine Learning algorithms. Furthermore, we list key challenges which pertain to the field in its current state. These challenges relate to difficulties in classification and fine-tuning machine learning models for use. Finally we provide some insight on the trajectory of the field and how it could potentially look in future years.

Keywords: Spectroscopy — Emission Lines — History of astronomy — Machine Learning — Support Vector Machine

1. INTRODUCTION

Within astronomy, one of the most astonishing fields is the analysis of spectral emission lines from galactic bodies. Spectral emissions are the bands of light which are uniquely radiated by atoms within galactic nebulae. More specifically, different atoms produce unique wavelengths of radiation. As a result, by studying emission lines, we are able to determine the atomic composition of Active Galactic Nuclei (AGNs) and star-forming regions which provides us with a myriad of information (SHIELDS 1999, 661). As such, this research focuses on classifying galaxies based on their emission spectrums, with a particular focus on distinguishing between Active Galactic Nuclei (AGNs) and star-forming regions.

Analyzing spectral emission line of galaxies, such as those arising from Hydrogen, Nitrogen, Oxygen, and Sulfur, into different categories brings immense value to the field of astronomy in that it provides crucial information about galaxies such as “chemical abundance, the amount of dust, the electron density, the age of the stellar population, the pressure of the interstellar medium, and the rate of star formation”. In particular, examining the ratios of emission lines of galaxies provides even greater substance to the field of astronomy since this allows for the mitigation of factors which interfere with measurements such as ISM pressure, red-shift effects, line bending, and error propagation (Kewley et al. 2019). Since this form of measurement requires a vast amount of data, using machine learning algorithms greatly increases the efficacy of its analysis.

There are a variety of methods to create machine learning algorithms to classify galaxies based on their emission spectrums. For our purposes, we have chosen to utilize K-means Clustering as an unsupervised machine learning model and Support Vector Machines (SVM) and Deep Learning Models as supervised learning models. The use of unsupervised and supervised models completed each other incredibly well as this allows us to compare new classification systems with previously established systems, which further allows us to weigh the impacts of the results of our k-means clustering model (Shi et al. 2015, 122). The use of these tools significantly improves our understanding of AGN’s and

star-forming regions which in turn provides critical data about galaxy formation as a whole, along with other equally important fields.

2. AN OVERVIEW OF EMISSION LINE GALAXIES

In the expanse of galactic research, the application of emission line spectra stands as a pivotal instrument, facilitating the nuanced classification of galaxies into discrete categories. To obtain a comprehensive understanding of how progress can be furthered in this field, it becomes essential to grasp the foundational aspects that have shaped its development. These foundational aspects encompass not only the scientific techniques and methodologies that have been refined over time but also the intellectual and conceptual shifts that have guided researchers towards new insights and provide us with a road-map of how we may continually advance this field of research in the near future.

The classification of galaxies is facilitated by the acquisition of spectroscopic data accompanied by flux measurements. This technique affords us the opportunity to derive significant insights into physical properties, such as metallicity and the ages of stars (Sánchez Almeida et al. 2012). In the context of the Sloan Digital Sky survey (SDSS), the methodology entails employing a 2.5 meter telescope equipped with spectroscopes that disperse light into constituent wavelengths, resolving an emission spectra. By quantifying spectra intensities through flux measurements, it becomes possible to draw elemental comparisons, leading to the acquisition of emission-line ratios, that provide insight into excitation properties that contribute to the classification of galaxies.

Within this context, the main areas of research within emission-line galaxy classification have been primarily concentrated on delineating between Active Galactic Nuclei (AGN) and regions primarily undergoing star-formation, two prominent sources of significant emission lines. The beginnings of AGN research can be traced to the work done by Carl Seyfert in 1943, where he investigated galactic emission lines. While analyzing spectrograms of six extra-galactic nebulae against a G-Type spectrum, Seyfert identified a distinct subset of these nebulae showcasing high excitation emission lines concentrated in their nuclei (Seyfert 1943). This breakthrough led to the discovery of Seyfert Galaxies, the first discovered class of AGNs. Typically positioned at the cores of galaxies, AGNs feature supermassive black holes that generate energy by accreting gas and dust and are extremely luminous. However, it wasn't until the advent of radio technology that AGNs gained significant prominence in research focus (Shields 1999).

The classifications of emission-line galaxies extends beyond the broad classes of AGNs and star-forming regions, encompassing further subcategories. While Seyfert's contributions unveiled Seyfert Galaxies, advancements in radio technology and subsequent research brought about additional sub categorizations of AGNs. These include Seyfert 1 and 2 classifications, LINERS (Low Ionization Nuclear Emission-Line Region Galaxies), and Quasars. Astronomers, Khachikian and Weedman contributed to the distinctions between Seyfert 1 and 2 Galaxies, where they were separated into groups based on the width of their emission lines (Khachikian & Weedman 1974). Seyfert 2 galaxies exhibit broader emission lines compared to Seyfert 1 galaxies. However, the widths of these lines at specific intensities typically range between 500 and 1000 km/s, which is narrower than the range observed in Seyfert 1 galaxies. This discrepancy in emission line width holds significance, as it suggests a divergence in the ionizing source between the two categories of galaxies (Khachikian & Weedman 1974). The discovery of LINERs can be attributed to T.M Heckman, publishing a paper on the subject in the 1980s. Heckman defined LINERs as objects with emission-lines from weakly ionized or neutral atoms such as O+, N+, O, and S and provided valuable insights into early-type galaxies, as LINERs were prevalent in the nuclei of nearly every galaxy of this type (Kauffmann 2009). An introduction to Quasars is provided in the paper authored by Jesse L. Greenstein and Maarten Schmidt in 1964. Studying the spectra of two quasi-stellar radio sources, 3C 48 & 3C 273, they were able to determine various properties of this new class of galaxies. (Greenstein & Schmidt 1964). Quasars, characterized by their great luminosity and emission spanning from ultraviolet to optical wavelengths, are known to primarily exhibit broad emission lines (Vanden Berk et al. 2001).

Conversely, another emission-line galaxy classification revolves around the characterization of star-forming regions, with a primary focus on HII regions. Approximately over the last two decades, our understanding of star-forming regions has expanded significantly, contributing to a deeper comprehension of galaxies and the early processes behind the formation of present-day galactic structures (Förster Schreiber & Wuyts 2020). One of the most common types of star-forming regions, HII regions, are defined as regions of star-formation where young stars emit large amounts of

energy leading to the ionization of adjacent interstellar gas, notably hydrogen, and therefore giving rise to distinctive emission lines (Peimbert, Peimbert & Delgado-Inglada 2017). The spectra of HII regions consist of a weak continuum, due to dust-scattered light, and strong emission lines. Distinguishing and grouping variations among emission-line spectra originating from HII regions and various sub-classes of AGN regions offers us insight into galactic evolution over cosmic time and the underlying physical properties that shape them.

3. EMISSION-LINE GALAXY CLASSIFICATION TECHNIQUES

The classification of emission-line galaxies into distinct groups begins with the development and utilization of different emission-line ratio diagrams, most notably the BPT (Baldwin-Phillips-Terlevich) Diagram, coinciding with the growing prominence of nebulae and active galactic nuclei study. In the 1981 paper published by Baldwin, Phillips, and Terlevich, they recognized the need for a new emission-line classification system that was more cohesive and based on excitation mechanisms with the growing increase of data on galaxy spectras (Baldwin, Phillips & Terlevich 1981). Previous efforts had involved the examination of galaxies' emission lines and had achieved a degree of differentiation (Smith 1975; Aller and Liller 1968). However, Baldwin, Phillips, and Terlevich aspired to formulate a straightforward two-dimensional classification framework aimed at enhancing the comparative analysis of distinct emission-line galaxies. To do so they compared the ratio of ($\lambda\lambda 3727/\lambda\lambda 5007$) against numerous other ratios such as ($\lambda\lambda 5007/\lambda\lambda 4861$), ($\lambda\lambda 6584/\lambda\lambda 6563$), ($\lambda\lambda 6300/\lambda\lambda 6563$), ($\lambda\lambda 4686/\lambda\lambda 4861$), ($\lambda\lambda 3426/\lambda\lambda 3727$), ($\lambda\lambda 3426/\lambda\lambda 5007$), and ($\lambda\lambda 6584/\lambda\lambda 5007$) with the purpose of distinguishing between gas clouds that were excited by shock-heating and power-law photo-ionization (Baldwin, Phillips & Terlevich 1981). In doing so, they were able to come up with four groups that could be created and separated, based on the dominant ionizing source that helped lead to the distinction between Star Forming regions, Seyfert Galaxies, and LINERs. Arguably the most widely-used version of the BPT diagram is the the N2 BPT diagram, one of their tested diagrams that plots the $[\text{OIII}]/\text{H}\beta$ vs $[\text{NII}]/\text{H}\alpha$ (Baldwin, Phillips & Terlevich 1981, Fig 5).

The efforts of Baldwin, Phillips, and Terlevich established the groundwork for emission-line ratio diagrams, which underwent ongoing expansions to enhance the classification of these galaxy types. In 1987, Sylvain Veilleux and Donald E. Osterbrock introduced another emission-line galaxy classification method that builds upon the BPT Diagram. This approach involves the utilization of two distinct diagrams: one compares the $[\text{OIII}]\lambda 5007/\text{H}\beta$ ratio against the $[\text{SII}]$ doublet ratio ($\lambda 6716/\lambda 6731$), and the other compares the $[\text{OIII}]\lambda 5007$ ratio against the $[\text{NII}]\lambda 6583/\text{H}\alpha$ ratio. These diagrams are accompanied by logarithmic equations, which serve the purpose of distinguishing various types of emission-line galaxies with an emphasis on separating narrow-line AGNs and HII region galaxies (Veilleux & Osterbrock 1987).

Astronomers Kewley, Groves, Kauffman, and Heckman also contributed to the field of study in their 2006 paper, "The Host Galaxies and Classification of Active Galactic Nuclei". Their work focused on separating AGN galaxies, Seyferts, LINERs, and Composites using optical emission-line ratios (Kewley et al. 2006). They found that $[\text{OIII}]/[\text{OII}]$ vs $[\text{OI}]/\text{H}\alpha$ diagnostic diagram very effectively separated Seyferts, LINERs, and galaxies dominated by star-formation, and used the distinctions made to interpret host properties of AGNs such as accretion rate and the ionization parameter (Kewley et al. 2006). Lamareille in 2009, shifted his focus to distinguishing between AGNs and Star-Forming regions at higher red-shifts. He used a diagram based on the emission-lines ratios of: $[\text{OIII}]\lambda 5007/\text{H}\beta$ and $[\text{OII}]\lambda\lambda 3726+3729/\text{H}\beta$ that lie in the blue part emission spectra (Lamareille 2009), terming it the blue diagram. He defined multiple different regions using logarithmic and quadratic boundary equations, effectively separating Seyfert 2, LINERs, composite galaxies, and star-forming galaxies. While Lamareille's diagram does have slight differences in categorizations when compared to the conventional red diagram like the BPT diagram, it offers greater insights into the examination of star-forming galaxies at elevated red-shifts (Lamareille 2009) and serves as an important classification tool.

Evidently, the advancements of diagnostic diagrams has been an ongoing process, serving as a productive means of classifying galaxies. While the BPT diagram remains relevant in the present time, progress in computational capabilities and the emergence of machine learning models led to new classification methods, especially considering the notable expansion of available data-sets. In 2015, astronomers from the North China Institute of Aerospace Engineering, developed a support-vector-machine model (SVM) to effectively classify galaxies as AGN hosts, star-forming

regions, or composite galaxies. A SVM closely resembles the diagnostic diagrams seeking the best hyperplane to separate classes, similar to how the BPT diagram uses logarithmic functions to separate classes. By utilizing input data including red-shift, color, and various emission-line ratios, they successfully trained and evaluated the SVM model, achieving a 98.9% accuracy in classifying the three distinct galaxy types (Shi et al. 2015). The researchers compared their SVM model with SVM models that replicated past diagnostic diagrams (using only two emission-line ratios as input) and found that using the entire set of inputs achieved better results. As the applications of machine learning continue to gain strength in research, its use in this domain is poised to further improve.

4. CURRENT CHALLENGES

Currently, there are a handful of key challenges persist in this field, one of which being the certainty of results. In the case where the result is more likely to be false, we would want our machine learning model output to reflect this. However, standard frequentist estimation paradigms for probability of that a result belongs in a cluster, tends to overestimate this probability as loss functions tend to bias the output towards certainty. In the case of support vector machine (SVM) classification, it is not entirely clear how to translate the distance from the separating hyperplane into a probability that the estimated classification is indeed the case. Having proper confidence in the model estimates, or uninterpretability of confidence in the case of SVM classification, can lead to more accurate insights being derived from the classification.

Another significant hurdle involves the automated identification and segmentation of the underlying components within every extensive classification cluster. As discussed earlier, there are a variety of sub-classes of Active Galactic Nuclei (AGN), such as Seyfert galaxies, which exhibit concentrated high emission lines (Seyfert 1943). If we are able to use unsupervised methods to automatically detect sub-clusters of each broader cluster, and in general use hierarchical clustering, then this will also give us useful insights on the workings of galaxies. For example, as mentioned earlier, the discovery of Quasars was made through observing a distinct cluster within the broader class of AGNs; that which contained broad emission lines from the optical to ultraviolet region (Greenstein & Schmidt 1964). Such breakthroughs could be aided by automatic classification of emission lines, were it to capture this hierarchical structure. A large limiting factor in determining the hierarchical structure, is simply the amount of data available to us. It would not be possible to determine hierarchical structure very efficiently if each sub-cluster had a very few number of galaxies belonging to the sub-cluster.

In addition, addressing the need for manual feature selection and transformation, is another prominent challenge within the field. The features being used are often nonlinear functions of emission line ratios, in order to transform the data as to be separated by a linear hyperplane. Currently, this process is done primarily by hand, through examination of the diagnostic diagrams such as in Lamareille in 2009 through quadratic nonlinearities. This analysis of the diagnostic diagrams by hand is also primarily how the emission lines that are used in the classification, are chosen. Ideally, this entire process of selecting and transforming the features should be automated as part of the classification pipeline itself. Selecting the proper emission lines to analyze will also heavily influence the clusters that result from unsupervised analysis of spectroscopic data, thereby being linked to a previously mentioned current challenge.

Furthermore, classifying composite galaxies poses another challenge. These galaxies have properties of both star forming and AGN galaxies; specifically, they have star forming properties in the optical part of their emission spectra while having AGN properties in the X-ray part of their emission spectra (Panessa et al. 2005), thus presenting a difficulty to current classification algorithms.

5. FUTURE DIRECTIONS

The fields of astronomy related to emission line spectra, and more broadly the entire study of astronomy and astrophysics, is expanding incredibly rapidly in the 21st century to unprecedented innovation in computational practices. This can most clearly be seen through the enormous rise in computing power through better hardware which allows for the flourishing of methods such as machine learning algorithms. This paper has gone into immense detail regarding the use of traditional and machine learning techniques which are incredibly useful in examining emission line ratios.

Even within machine learning, there are a plethora of other algorithms which could lead to breakthroughs within the analysis of spectral emission lines. As such, there are numerous routes in which this field could develop, along with differing algorithms, this classification can be further refined to differentiate between sub-fields within AGN and star-forming regions such as quasars and supernova.

One of the ways in which emission line research can further develop is through advanced machine learning techniques, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) which serve as potentially create higher accuracy galaxy classification models than is possible with K-means clustering and SVM. CNNs are particularly unique in that it “uses convolutional layers in place of the majority (or all) of the dense layers”. As such, CNNs can be trained to use the optimal layers to perform tasks such as galaxy classification. Similarly RNNs are incredibly useful in enhancing galaxy classification techniques. This can be attributed to its property of many-to-one encoding and its ability to be expanded upon into gated recurrent neural networks (Smith & Geach 2023).

These advanced machine learning techniques are incredibly useful in evaluating the capability of galaxy classifications based on emission spectrum lines. As such, this sets the foundation for greater precision classification which could differentiate between sub-fields such as quasars, blazars, supernova, and various galactic nebulae. For example, previous research has indicated that radio-loud AGN, blazars and radio galaxies can be classified through the use of [OIII] and [OII] emission lines (Landt et al. 2004, 83). Through the use of advanced machine learning algorithms such as CNNs and RNNs, research regarding classification of galaxy types through emission line spectra could advance to the point where the use of these algorithms could be the default technique for differentiating them. This could even potentially largely resolve the challenge of classifying composite galaxies as discussed previously. Naturally, with these developments in precision and accuracy, the scale of this field could greatly increase to the point where these techniques could be embedded within major data collecting surveys.

6. SUMMARY AND OUTLOOK

In conclusion, this survey highlights some of the key insights and offers a comprehensive overview of the field of emission-line galaxy classification, spanning its historical origins to the contemporary integration of cutting-edge machine learning techniques, and the future direction the field looks to take. Beginning with Carl Seyfert’s groundbreaking discoveries in the mid-1900s, the field’s evolution has been propelled by advancements in spectroscopy and radio technology. This progress led to further advancements in the identification of distinct galaxy sub-classes, including LINERs and Quasars and the distinction between Seyfert 1 and 2 galaxies. Diagnostic diagrams such as the renowned BPT diagram and further extensions on it have played a pivotal role in classifying galaxies by considering ionization mechanisms, particularly those arising from AGN host galaxies and star-forming regions and are still widely used to this day.

With the advent of extensive data-sets, such as the Sloan Digital Sky Survey (SDSS), and advancements in computation, a new approach in machine learning has been adopted to process and analyze vast amounts of spectral data efficiently. While certain challenges loom, such as the need to establish a robust hierarchical classification structure, enhance result confidence levels, and refine feature selection and manipulation techniques, we remain committed to overcoming these hurdles. Our primary objective is to develop both unsupervised and supervised machine learning models, trained on the abundant amount of spectral information from the SDSS. In doing so, we hope to present the scientific community with a more precise and efficient tool for the categorization of emission-line galaxies, with the ultimate goal of reshaping our fundamental understanding of these cosmic entities and their role in the broader context of galaxy evolution.

REFERENCES

- Baldwin, J. A., Phillips, M. M., & Terlevich, R. 1981, *PASP*, 93, 5
- Förster Schreiber, N. M., & Wuyts, S. 2020, *Annu Rev Astron Astrophys*, 58, 661
- Greenstein, J. L., & Schmidt, M. 1964, *ApJ*, 140, 1
- Kauffmann, G. 2009, *A& A*, 500, 201
- Kewley, L. J., Groves, B., Kauffmann, G., & Heckman, T. 2006, *Monthly Notices of the Royal Astronomical Society*, 372, 961
- Kewley, L. J., Nicholls, D. C., & Sutherland, R. S. 2019, *Annu Rev Astron Astrophys*, 57, 511
- Khachikian, E. Y., & Weedman, D. W. 1974, *ApJ*, 192, 581
- Lamareille, F. 2010, *A& A*, 509, A53
- Landt, H., Padovani, P., Perlman, E. S., & Giommi, P. 2004, *Monthly Notices of the Royal Astronomical Society*, 351, 83
- Panessa, F., Wolter, A., Pellegrini, S., et al. 2005, *ApJ*, 631, 707
- Peimbert, M., Peimbert, A., & Delgado-Inglada, G. 2017, *PASP*, 129, 082001
- Sanchez Almeida, J., Terlevich, R., Terlevich, E., Cid Fernandes, R., & Morales-Luis, A. B. 2012, *ApJ*, 756, 163
- Seyfert, C. K. 1943, *ApJ*, 97, 28
- Shields, G. A. 1999, *PUBL ASTRON SOC PAC*, 111, 661
- Shi, F., Liu, Y.-Y., Sun, G.-L., et al. 2015, *Mon Not R Astron Soc*, 453, 122
- Smith, M. J., & Geach, J. E. 2023, *R Soc open sci*, 10, 221454
- Veilleux, S., & Osterbrock, D. E. 1987, *ApJS*, 63, 295
- Vanden Berk, D. E., Richards, G. T., Bauer, A., et al. 2001, *The Astronomical Journal*, 122, 549