Multidimensional Data Driven Classification of Active Galaxies

Vasileios Stampoulis
Supervisor : Prof. David van Dyk

Department of Mathematics
Imperial College London

London, October 2016

Outline

1. The Classification Problem
   - The Scientific Problem
   - The Dataset
   - Existing Classification methods
   - Inefficiency of Existing Methods
2. Clustering Methodology
   - Introduction
   - Finite Mixture Models
   - Stylized Facts of the Dataset
   - Proposed Approach
   - Comparing the Classification Schemes
3. Multidimensional Decision Boundaries
4. Discussion

Outline

1. The Classification Problem
   - The Scientific Problem
   - The Dataset
   - Existing Classification methods
   - Inefficiency of Existing Methods
2. Clustering Methodology
   - Introduction
   - Finite Mixture Models
   - Stylized Facts of the Dataset
   - Proposed Approach
   - Comparing the Classification Schemes
3. Multidimensional Decision Boundaries
4. Discussion

The Scientific Problem

- Spectroscopy is the study of the measurement of radiation intensity as a function of wavelength.
- Spectroscopy has been utilised in identifying the main power source in active galaxies.

The Scientific Problem

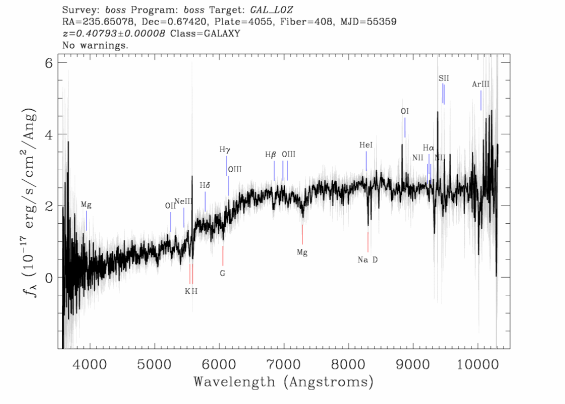

Figure: A selected spectrum from the DR10 BOSS data, showing absorption (red) and emission (blue) lines (http://www.sdss.org).

The Scientific Problem

Based on the different mechanisms that excite the gas inside the galaxies (and which, as a result of those mechanisms, glows in different wavelengths), the Galaxies may be separated, according to Kewley et al (2001):

- H_{II} region-like galaxies (star forming galaxies, SFG).
- AGN (Active Galactic Nuclei) which can be divided further to LINER and Seyferts.
- Composite galaxies which exhibit characteristics from both AGNs and SFGs.

The Dataset

- Sloan Digital Sky Survey (SDSS).
- The SDSS is a major multi-filter imaging and spectroscopic redshift survey.
- The latest Data Release (DR 10) the from SDSS-IIIs Baryon Oscillation Spectroscopic Survey (BOSS) contains 1,848,851 Optical Galaxy spectra.

Existing Classification methods

- From early on (see Alloin et al. (1978)), intensity ratios, such as $O_{III}/H\beta$, have been utilised as means to classify the emission line spectra of galactic H_{II} region.
- Baldwin et al. (1981) proposed three diagnostic diagrams based on four emissions intensities ratios; $\log(N_{II}/H\alpha)$, $\log(S_{II}/H\alpha)$, $\log(O_{I}/H\alpha)$ and $\log(O_{III}/H\beta)$.
- Those three diagrams, known as **Baldwin-Phillips-Terlevich (BPT)** diagrams, were introduced in order to classify the dominant energy source in emission-line galaxies.

Existing Classification methods

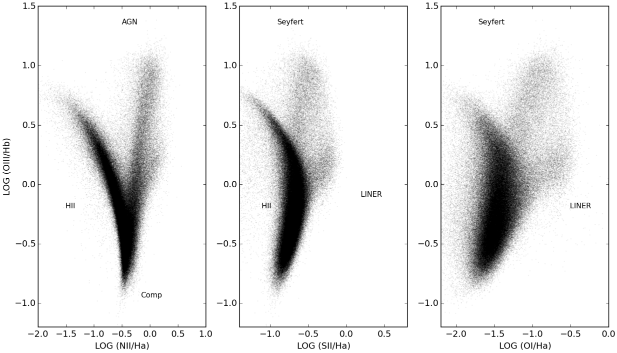

Figure: Example diagnostic (BPT) diagram based on observational data from the DR10 BOSS data.

Existing Classification methods

In the **first diagnostic**:

- There are two different loci of sources; a stream moving from the bottom middle to the top left and another fuzzier stream moving from the bottom middle to the top right.
- The first stream corresponds to the star-forming galaxies and the latter to the AGN.

The **second and third diagnostics** verify that the AGN are made up of two groups, the Seyfert and the LINER.

Existing Classification methods

- Kewley et al. (2001) introduced a maximum ’starburst’ line on the BPT diagrams.
- This line (red line in the next slide) defines the upper limit of the star-forming galaxies and was developed using theoretical models.

- Kauffmann et al. (2003) presented an empirical line to distinguish the pure SFGs from Seyfert- SFG **composite objects** (cyan dashed line in the next slide).
- The spectra of those composite galaxies contain significant contributions from both AGN and star-formation according to Kewley et al. (2006).

Existing Classification methods

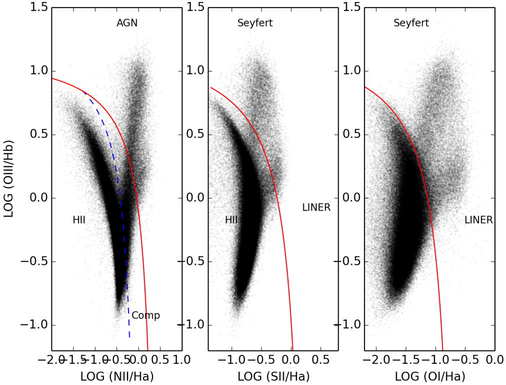

Figure: Diagnostic (BPT) diagram based on observational data from the DR10 BOSS data with the maximum 'starburst' line (red) and composite line (cyan) plotted.

Inefficiency of Existing Methods

- The theoretical model of Kewley et al. (2001) fails to distinguish the AGN from the star-forming galaxies.
- The use of the three diagnostics independently of each other often gives contradicting classification for the same source.
- In Kewley et al. (2006), 8 percent of the galaxies in their sample are characterized as ambiguous galaxies.

We propose a new soft clustering scheme, the soft allocation data driven method (SoDDA), for classifying galaxies using emission-line ratios.

Outline

1. The Classification Problem
   - The Scientific Problem
   - The Dataset
   - Existing Classification methods
   - Inefficiency of Existing Methods
2. Clustering Methodology
   - Introduction
   - Finite Mixture Models
   - Stylized Facts of the Dataset
   - Proposed Approach
   - Comparing the Classification Schemes
3. Multidimensional Decision Boundaries
4. Discussion

Introduction

- Cluster analysis is a statistical method that aims to partition a dataset into subgroups so that the members within each subgroup are more homogeneous (according to some criterion) than the population as a whole.
- The Probabilistic (model-based) algorithms assume that the data is an i.i.d. sample from a population described by a density function, which is taken to be a mixture of component density functions.
- It is common to assume that the mixture components are all from the same parametric family, such as the normal.

Finite Mixture Models

Given data x ($n \times p$) with independent multivariate observations $x_1, ..., x_n$, the joint density for a Finite Mixture model with K components:

$$p(x|\theta, \pi) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f_k(x_i|\theta_k), \tag{1}$$

where f_k and θ_k are the density and parameters of the component k in the mixture model and π_k is the probability that an observation comes from the component k. Furthermore, $\pi_k \geq 0$ and $\sum_{i=1}^{K} \pi_k = 1$.

EM Algorithm

Dempster et al. (1977) proposed a framework for computing the maximum likelihood estimators in finite mixture models for a general specification of the component density functions f_k using the **Expectation-Maximization (EM) algorithm**.

- The EM algorithm is an iterative method for posterior mode finding.
- Given a statistical model consisting of a set x of observed data, a set of unobserved latent data or missing values z, and a vector of unknown parameters θ, the EM alternates between performing an expectation (E) step and a maximization (M) step.

Finite Mixture Models cont.

In the context of **Finite Mixture models** Dempster et al. (1977):

- Introduced an unobserved vector z $(n \times K)$, where z_i is the indicator vector of length K with $z_{ik} = 1$ if the observation x_i belongs to cluster k and 0 otherwise.
- The z's are latent variables that specify from which component each observation is drawn.

Finite Mixture Models cont.

The joint distribution $p(x, z|\theta, \pi) = p(z|\theta, \pi)p(x|z, \theta, \pi)$.

- $p(z|\theta, \pi)$ is multinomial distribution $p(z|\theta, \pi) = \prod_{k=1}^{K} \pi_k^{z_k}$, $\theta = (\theta_1, ..., \theta_K)$ and $\pi = (\pi_1, ..., \pi_K)$.
- Conditional on $z_{ik} = 1$, $x_i \sim f_k(x_i|z_{ik} = 1, \theta_k)$.

The resulting complete-data log-likelihood (the likelihood of both the observed and missing data) is:

$$\ell(\theta, \pi|x, z) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} log[\pi_k f_k(x_i|\theta_k)]$$

Multivariate Gaussian Mixture Model

- The Multivariate Gaussian mixture model has been used with success for many clustering problems.
- Multivariate normal mixtures can be used for data with varying structures due to the flexibility in how we define the covariance matrices.

The Multivariate Normal distribution is defined as:

$$f_x(x_1, ..., x_p) = \frac{1}{\sqrt{(2\pi)^p|\Sigma|}} exp\Big(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\Big)$$

Limitations of the Multivariate Normal Model

- Fraley et al. (2002) pointed out that data generated by mixtures of multivariate normal distributions are characterized by clusters centered at the means μ_k with increased density for points closer to the means.
- The practical use of multivariate mixture models is limited for data that exhibit non-normal features, including asymmetry, multi-modality and heavy tails as in the SDSS dataset.

Non Gaussian Features of the Dataset

- The distribution of the Star forming Galaxies (the HII-region) is clearly highly skewed and not-linear (curve-shaped).
- In the LINER and Seyfert clusters, the mass doesn't seem to be concetrated at the centers.
- There seems to be a big overlap between the 4 different clusters.

Fitting *K* Multivariate Normal Distributions

- We use a mixture of MG distributions with *K* considerably larger than the actual number of galaxy classes.
- This allows a great deal of flexibility in the class-specific distributions of emission line ratios.
- Then we perform hyper-clustering of the *K* MG distributions so as to concatenate them into subpopulations representing the four desired galaxy classes.

Fitting *K* Multivariate Normal Distributions

- The number ($K >> 4$) of MG distributions that we fit to our data is chosen using the Bayesian Information Criterion (BIC)

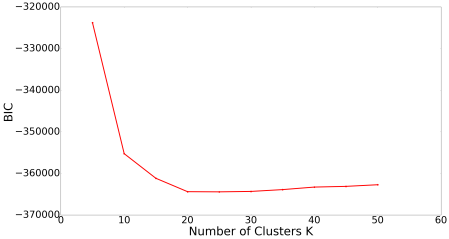

Figure: The Bayesian Information Criterion (BIC) computed over a grid of values of *K* (using increments of 5) using the data of the SDSS DR 10.

Fitting 20 Multivariate Normal Distributions

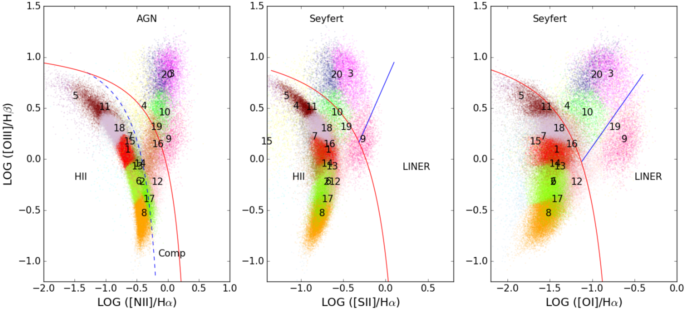

Figure: The BPT diagnostic diagrams for the SDSS DR10 sample; each datapoint is coloured according to its most probable allocation to one of the 20 multivariate Gaussian Distributions.

Combining the 20 Multivariate Normal Distributions

- SoDDA accomplishes the hyper-clustering of the *K* subpopulations into the four galaxy classes using the classification scheme of Kewley et al. (2006).
- We treat the fitted subpopulations means $(\mu_1^\star, ..., \mu_K^\star)$ as a dataset and classify them into the four galaxy classes.
- Thus, we define the distribution of the emission line ratios for each galaxy class as a finite mixture of MG distributions.

Combining the 20 Multivariate Normal Distributions

Table: The suggested classification of the 19 subpopulations means.

Class	Subpopulation ID
SFG	1,2,5,6,7,8,11,13,14,15,17,18
Seyferts	3,10,20
LINER	9
Composites	12,16,19

We opted to discard cluster 4. It is located in a different region on each one of the three diagnostic diagrams and inspection of the optical spectra of several of the sources allocated to this cluster, shows that they have very broad emission lines with complex structure, resulting in unreadable line measurements.

Combining the 20 Multivariate Normal Distributions

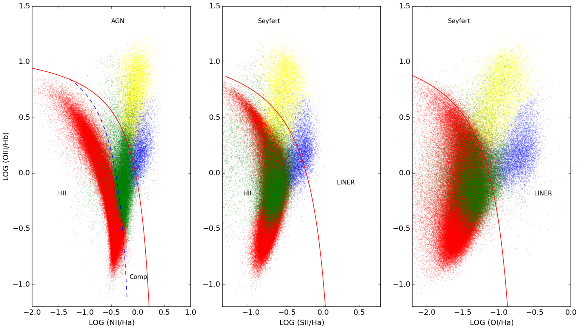

Figure: Combining the 20 mutltivariate Normals in the SDSS dataset as explained above and plotting each datapoint with different color according to the most probable allocation.

Combining the 20 Multivariate Normal Distributions

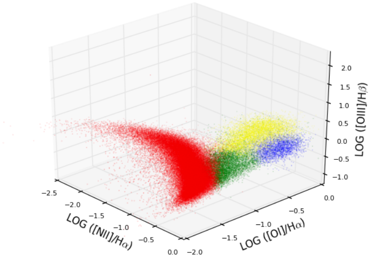

Figure: A 3-dimensional projection of the SDSS DR10 sample on the ($\log(N_{II}/H\alpha)$, $\log(S_{II}/H\alpha)$, and $\log(O_{III}/H\beta)$) volume, in which each datapoint is plotted with different colour according to the allocation from SoDDA.

The theoretical Classification

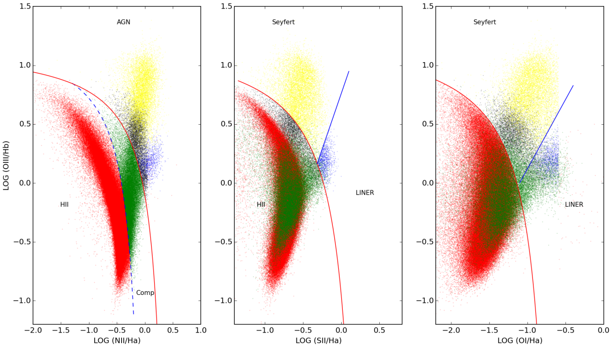

Figure: The SDSS dataset in which each datapoint is plotted with different color according to the classification scheme of Kewley et al. (2006).

Comparing the Classification Schemes

Table: A 3-way classification table that compares the SoDDA classification with the standard, 2-dimensional classification scheme Kewley et al. (2006). Each cell has 3 values: the number of galaxies with (i) $\rho_{ic} \geq 75\%$, (ii) $50\% \leq \rho_{ic} < 75\%$, and (iii) $\rho_{ic} < 50\%$, where ρ_{ic} is the posterior probability that galaxy *i* belongs to galaxy class *c* under SoDDA.

		Kewley et al. (2006)																	
		SFGs			Seyferts			LINERs			Comp			Ambiguous			Total		
SoDDA	SFGs	65080	946	3	6	0	0	0	0	0	927	1343	69	1744	62	8	67757	2351	80
	Seyferts	0	0	0	5471	262	8	0	0	0	2	28	10	349	1131	45	5822	1421	63
	LINERs	0	0	0	9	33	5	778	4	0	700	181	15	891	234	44	2378	452	64
	Comp	57	251	4	32	40	4	0	0	0	4211	2668	103	258	801	38	4558	3760	149

Outline

1 The Classification Problem
- The Scientific Problem
- The Dataset
- Existing Classification methods
- Inefficiency of Existing Methods

2 Clustering Methodology
- Introduction
- Finite Mixture Models
- Stylized Facts of the Dataset
- Proposed Approach
- Comparing the Classification Schemes

3 Multidimensional Decision Boundaries

4 Discussion

4-dimensional Linear Decision Boundaries

- The multidimensional decision boundaries implied by SoDDA can be obtained using support vector machine (SVM).
- A SVM is a discriminative classifier formally defined by a separating hyperplane.
- Given classified galaxies, the algorithm outputs an optimal hyperplane which can be used to categorise new unlabelled galaxies.

4-dimensional Linear Decision Boundaries

Table: Comparing the classifications of SoDDA with that of the 4-dimensional SVM.

		SoDDA				
		SFGs	Seyferts	LINERs	Composites	Total
SVM	SFGs	69794	0	0	573	70367
	Seyferts	3	7033	59	273	7368
	LINERs	0	20	2703	143	2866
	Composites	391	253	132	7478	8254
	Total	70188	7306	2894	8467	

3-dimensional Linear Decision Boundaries

- The [O_{I}] line is generally hard to observe.
- Thus, we fit the SVM algorithm to the SDSS DR10 dataset using the classifications from SoDDA and only the 3 emission line ratios log(N_{II}/Hα), log(S_{II}/Hα), and log(O_{III}/Hβ)

3-dimensional Linear Decision Boundaries

Table: Comparing the classifications of SoDDA with that of the 3-dimensional SVM.

		SoDDA				
		SFGs	Seyferts	LINERs	Composites	Total
SVM	SFGs	69746	0	7	808	70561
	Seyferts	5	7010	99	278	7392
	LINERs	0	66	2574	154	2794
	Composites	437	230	214	7227	8108
	Total	70188	7306	2894	8467	

Outline

1. The Classification Problem
   - The Scientific Problem
   - The Dataset
   - Existing Classification methods
   - Inefficiency of Existing Methods
2. Clustering Methodology
   - Introduction
   - Finite Mixture Models
   - Stylized Facts of the Dataset
   - Proposed Approach
   - Comparing the Classification Schemes
3. Multidimensional Decision Boundaries
4. Discussion

Discussion

- We propose a new soft clustering scheme, the soft allocation data driven method (SoDDA), for classifying galaxies using emission- line ratios.
- The main advantages of this method are the use of all four optical-line ratios simultaneously.

Further Research Direction

- Extend to include **additional diagnostic lines** (e.g. OII or IR lines PAH, OIV, NeII) information in other wavamebds (e.g combination of optical line ratios with IR colours such as the Stern or Donley diagnostics, or flux ratios in different wavelengths (e.g. X-ray to optical ratio).
- Use SoDDA in order to create $\log N - \log S$ **diagrams** for AGNs and star-forming Galaxies.

Thank you!

Many Thanks to:

David van Dyk
Andreas Zezas
Vinay Kashyap

Bibliography

- Alloin, D., Bergeron, J., & Pelat, D. (1978). Properties of a Sample of Irregular Galaxies. Astronomy and Astrophysics, 70, 141.
- Baldwin, J. A., Phillips, M. M., & Terlevich, R. (1981). Classification parameters for the emission-line spectra of extragalactic objects. Publications of the Astronomical Society of the Pacific, 5-19.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 1-38.
- Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association, 97(458), 611-631.
- Kauffmann, G., Heckman, T. M., Tremonti, C., Brinchmann, J., Charlot, S., White, S. D., ... & Schneider, D. P. (2003). The host galaxies of active galactic nuclei. Monthly Notices of the Royal Astronomical Society, 346(4), 1055-1077.