Machine Learning in Astronomy: Possibilities and Pitfalls
Proceedings IAU Symposium No. 368, 2025
J. McIver, A. Mahabal & C. Fluke, eds.
doi:10.1017/S1743921323000728

Unsupervised classification: a necessary step for Deep Learning?

Didier Fraix-Burnet

Univ. Grenoble-Alpes / CNRS / IPAG, France
email: didier.fraix-burnet@univ-grenoble-alpes.fr

Abstract. Deep-learning algorithms have gained much popularity in the past decade. However, their supervised nature makes them fully dependent on the quality and completeness of the data samples used in the training processes. And when it comes to galaxy spectra for instance, there is simply no suited training sample available yet. This is where unsupervised classification (clustering) can come to the rescue.

Using the discriminant latent mixture-model based algorithm Fisher-EM, we 1) demonstrate the discriminative capacity of the method and its robustness with respect to noise on a sample of galaxy spectra simulated with the code CIGALE, 2) manage to classify ∼700 000 spectra of galaxies from the SDSS, and 3) extend this classification to higher redshifts thanks to the VIPERS data (∼80 000 spectra up to a redshift of 1.2). Finally, it appears that FisherEM is efficient for images of galaxies as well.

Keywords. Astrostatistics, Galaxies, Spectra, Unsupervised classification

1. Introduction

Classification of galaxies is often synonymous with the Hubble's morphological classification. The goal of modern extragalactic astronomy is to automatically classify billions of images in a very short time. Deep Learning sophisticated techniques, a priori perfectly suited for this task, rely on training samples that are based on this classification. Here comes a paradox: the current data are not only enormous in number but they are very much detailed, while the classes made by Hubble were based on a very limited information. In addition, the Hubble morphological classification is made by eye, a very lengthy and subjective procedure. The slightly more sophisticated classification by de Vaucouleurs (18 classes) is rarely used for the labelling of the training samples despite the fact that it corresponds better to modern detailed images. But this classification requires a very careful human examination with a high level of expertise, that Deep Learning is unable to reproduce.

We believe that our incredible data deserve better suited classification schemes, and the logical approach is to remain data-driven throughout the process. Noting that the Hubble classification was fully unsupervised, we promote the use of clustering (unsupervised classification) to describe the output of modern telescopes more faithfully and feed the supervised Artificial Intelligence techniques with more consistent labellings.

There is no global classification for the spectra of galaxies, apart from some taxonomy based on specific features. Somehow this is a chance to build something from scratch using unsupervised approaches. Our group is the first to propose an entirely data-driven atlas of galaxy spectra obtained from the whole optical spectra.

© The Author(s), 2025. Published by Cambridge University Press on behalf of International Astronomical Union.

https://doi.org/10.1017/S1743921323000728 Published online by Cambridge University Press

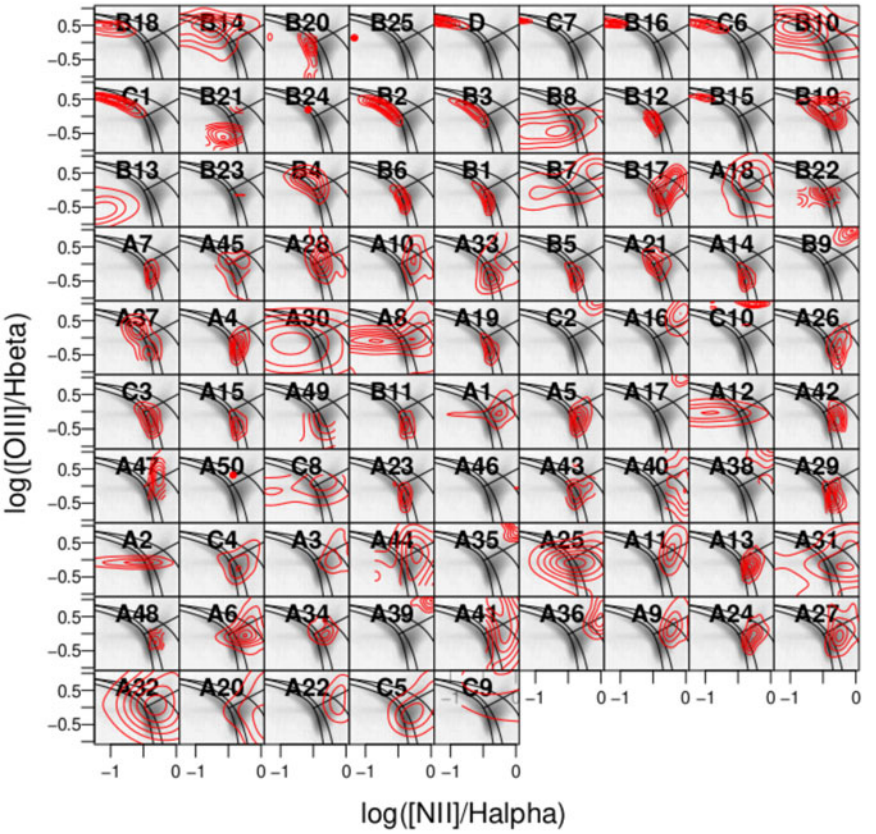


Figure 1. BPT diagram for the 86 classes of the SDSS samples of galaxy spectra ordered according to the slope of the mean spectra of each class.

2. Method: the algorithm FisherEM

A basic clustering technique is the k-means algorithm which tries to fit a mixture of identical Gaussians to a data set. Developments of this approach have introduced the possibility for the classes to have different distributions through the well-known optimisation procedure called expectation-maximisation (EM) iterations. Latent subspaces constitute further improvements for the clustering through a reduction of the dimensionality of the data set. The algorithm we use (FisherEM†; Bouveyron & Brunet 2012; Bouveyron & Brunet-Saumard 2014; available in R) introduces a supplementary step in the EM iteration to optimise the latent subspace based on the Fisher criterion. This Fisher criterion is the ratio between the variance between the classes to the variance within the classes. This guarantees that the latent subspace is optimised for the clustering goal.

3. Physical relevance of the classes

From a set of simulated spectra of galaxies obtained with the code CIGALE (Dubois et al. 2022), we have shown that each class corresponds to specific physical properties. Moreover, we show that the classification is robust with respect to the level of signal to noise ratio, up to three. But even at a signal to noise ratio of one, the obtained

† https://www.rdocumentation.org/packages/FisherEM/versions/1.6

https://doi.org/10.1017/S1743921323000728 Published online by Cambridge University Press

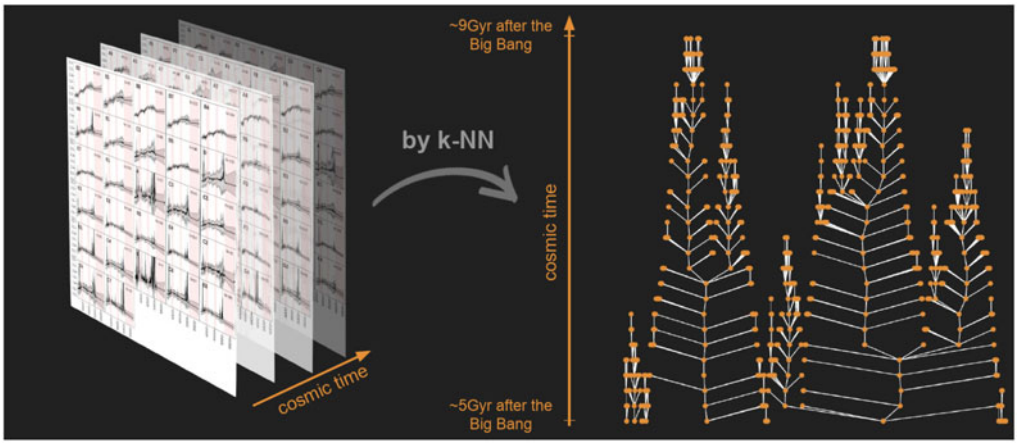


Figure 2. Tree of classes of galaxy spectra for the VIPERS sample. Courtesy Julien Dubois.

classification is consistent despite being less detailed. This result gives a lot of confidence in the capabilities of the algorithm FisherEM.

4. An atlas of SDSS galaxy spectra

We performed an unsupervised classification of 702248 spectra of galaxies and quasars with redshifts smaller than 0.25 that were retrieved from the Sloan Digital Sky Survey (SDSS) database, release 7 (Fraix-Burnet et al. 2021). The optimum number of classes given by a penalised likelihood criterion called Integrated Complete Likelihood (ICL) is 86 classes, of which the 37 most populated ones gather 99% of the sample.

A preliminary physical interpretation of the classes (Fraix-Burnet et al. 2021) show that the spectra are well discriminated by their slope, but more remarkably by their emission line ratios as illustrated by the BPT diagram of [OIII]/H_β versus [NII]/H_α (Fig. 1).

This is the first time that an automatic, objective and robust unsupervised classification is established on such a large number of galaxy spectra. We are currently working on the detailed interpretation of each of the classes in order to provide a physical atlas of galaxy spectra at low redshift.

5. Extending the classification up to redshift 1.2

Going to higher redshifts is difficult because the information contained in the spectra varies. The only solution is to divide the samples into bins of redshift, compromising between the number of spectra and the number of common data points. We have succeeded in obtaining an unsupervised classification for each of the 26 bins between redshifts 0.4 to 1.2 on a sample of 80 000 spectra from the VIPERS survey†. We then used the k-Nearest Neighbour (kNN) technique to connect the classes between two consecutive bins to provide an "evolutionary" tree (Fig. 2, Dubois et al., in prep).

6. Conclusion

Our results can be used in several ways. Firstly, we provide a data-driven atlas of galaxy spectra with its evolution in redshift. Secondly, the statistical mixture-based model can be used to instantaneously classify any new spectra by simply computing the probabilities

† http://www.vipers.inaf.it/

https://doi.org/10.1017/S1743921323000728 Published online by Cambridge University Press

for this spectrum among all the classes. Lastly, our classification makes a data-driven training sample of nearly 800 000 optical spectra of galaxies up to a redshift of 1.2 that can be used for instance in Deep Learning approaches.

We finally note that we have obtained excellent results on images of galaxies with this same algorithm FisherEM.

References

Bouveyron, C., Brunet, C. 2012, *Statistics and Computing*, 22, 301
Bouveyron, C., Brunet-Saumard, C. 2014, *Computational Statistics*, 29, 489
Dubois, J., Fraix-Burnet, D., Moultaka, J., *et al.* 2022, *A&A*, 663, A21
Fraix-Burnet, D., Bouveyron, C., Moultaka, J. 2021, *A&A*, 649, A53

https://doi.org/10.1017/S1743921323000728 Published online by Cambridge University Press