Photometric AGN Classification in the SDSS

S. Cavuoti[1,2,3], R. D'Abrusco[1,2,3], G. D'Angelo[1,2,3], M. Brescia[3,4], N.V. Deniskina[1,2,3], O. Laurino[1], & G. Longo[1,2,3,4]

1 *- Università degli Studi di Napoli Federico II;* ***2*** *- PON SCoPE;* ***3*** *- INAF, Istituto Nazionale di Astrofisica;* ***4*** *INFN, Napoli Unit*

We present the results of a supervised neural network approach to the problem of the search for candidates Active Galactic Nuclei in photometric data using a spectroscopic base of knowledge. Due to the computational costs, we used Multi Layer Perceptron and Support Vector Machine algorithms run on the SCOPE, COMETA & CYBERSAR GRID The results are far better than those obtained with more traditional methods.

Our work aims at obtain a way to classify AGN using photometric data instead of spectra. AGN selection is usually performed from the overall spectral distribution, together with some spectroscopic indicators like equivalent linewidth, FWHM of specific lines or lines flux ratios. A reliable and accurate AGN classificator based on photometric features only, would allow to save precious telescope time and enable several studies based on statistically significant samples of objects.

AGN classification is made usually by means of some diagnostic diagrams (usually called BPT) that involve some emission line ratios, and/or FWHM. In this diagrams AGN and not-AGN are empirically separated by some lines. We used the Kewley, Kauffman and Heckman lines.

$$\log\frac{[OIII]\lambda5007}{H_\beta} = \frac{0.61}{\log\frac{[NII]\lambda6583}{H_\alpha} - 0.47} + 1.19$$

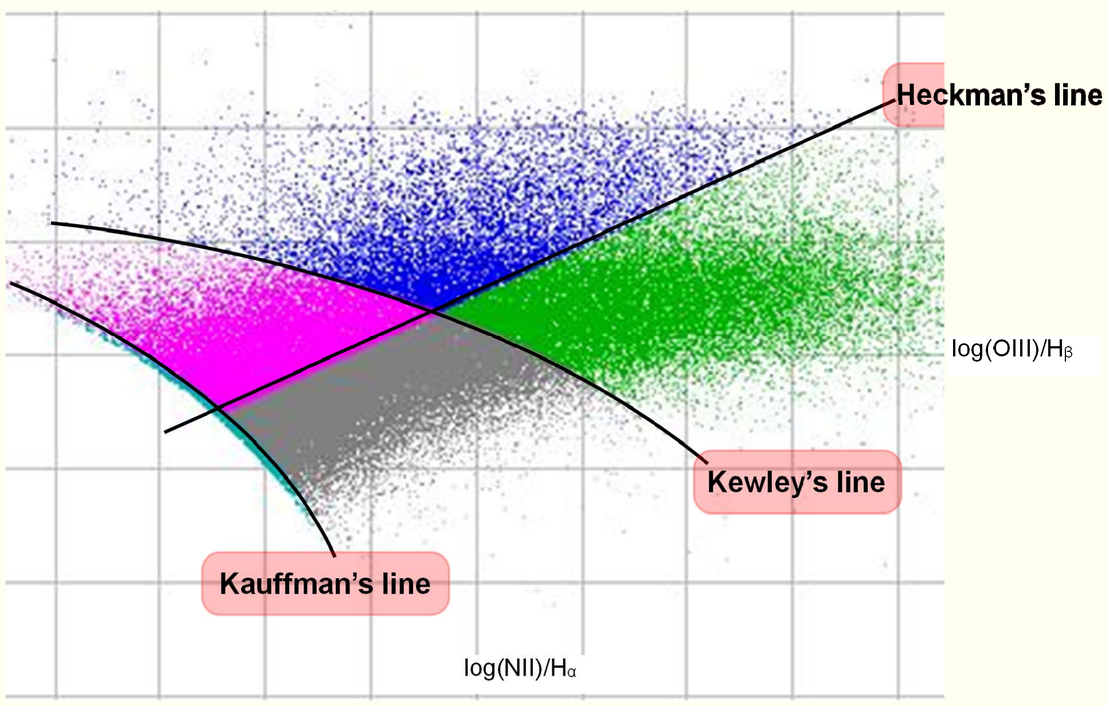
Kewley line L.J. Kewley et al.: ApJ, 556, 121 (2001)

$$\log\frac{[OIII]\lambda5007}{H_\beta} = \frac{0.61}{\log\frac{[NII]\lambda6583}{H_\alpha} - 0.05} + 1.3$$

Kauffman line G. Kauffman et al.: The host galaxies of active galactic nuclei, MNRAS, 346, 1055, (2003)

$$\frac{[OIII]\lambda5007}{H_\beta} = 2.1445\frac{[NII]\lambda6583}{H_\alpha} + 0.465$$

Heckman line T.M. Heckman: A&A, 87, 182 (H80) (1980)



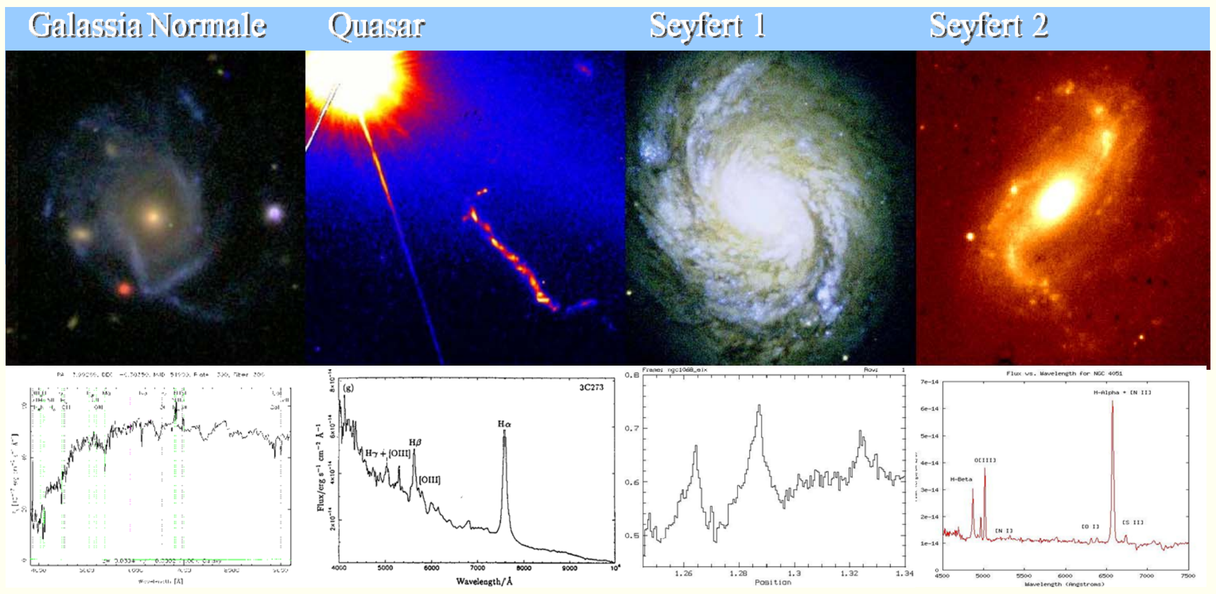
The knowledge base (BoK) used to create the target vectors is crucial since neural networks have no power of extrapolations, and if the spectroscopic targets do not cover the whole photometric parameter space, all the biases are reproduced by the nerual networks. Our BoK was obtained by joining three catalogues:

- A catalogue made by Sorrentino et al. (2006), that separates (0.05<z<0.095) object into Seyfert 1, Seyfert 2 and Not AGN. (an object is considered to be an AGN if it falls above one of the Kewley's lines. Once one object is declared as AGN is declared to be a Seyfert 1 if: FWHM(H_α) > 1.5 FWHM ([OIII]λ 5007) or FWHM(H_α) >1200 Km/s; all other AGN are classified as Seyfert 2.
- A catalogue contained spectra lines and ratio for 88178 galaxies (0.02<z<0.3) and, according to the work of Kauffman et al. (2003), we define a zone where there are just AGN that is above the Kewley's line; a zone where objects are not AGN, below the Kauffman's line, and a mix zone where AGN and not AGN overlap. The Mix and Pure AGN zone are divided into Seyfert and LINERs zone by the Heckman's line.
- A catalogue made by D'Abrusco et al. (2007) and contains the photometric redshift (with an accuracy estimated by σ_{rob}= 0.02).

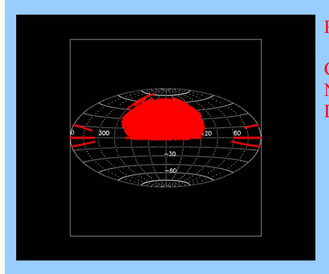
We made three experiments with MLP and SVM, and for all of them we used the same features: *petroR50_u, petroR50_g, petroR50_r, petroR50_i, petroR50_z, concentration_index_r, z_phot_corr, fibermag_r, (u – g) dered, (g – r) dered, (r – i) dered, (i – z) dered, dered_r.* All features, with the exception of *z_phot_corr* (from D'Abrusco's catalogue) were taken from the SDSS database.

We perform this experiments:

(1)"AGN vs Mix"
(2)"Type1 vs Type2"
(3)"Seyfert vs LINER"



DATA: SLOAN DIGITAL SKY SURVEY (SDSS)

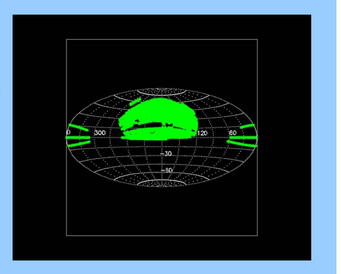


Photometry:

Coverage: 9583 sq. deg.
Number of Object: 287 M
Data Volume:
Images: 10 TB
Catalogs: 6 TB

Spectroscopy:

Coverage: 7425 sq. deg.
Number of Object: 1.2 M
Data Volume: 310 GB



We need an automatic approach to data analysis (data mining) and great computing power (GRID)

DATA MINING INSTRUMENTS:

Multi Layer Perceptron

The algorithm known as Multi Layer Perceptron (MLP) is based on the concept of perceptron, derived from the biological neuron, while the method of learning is based on gradient-descent method that allows you to find a local minimum of a function in a space with N dimensions. The weights associated to the connections between the layers of neurons, initialized at small and random values, and then the MLP applies the learning rule using the template patterns.

Support Vector Machines

Given the training vectors $x_i \in R^n$, i = 1...l, in two classes, and a vector $y \in R^l$ such that y_i {1,-1}, C-SVC (Boser et al., 1992; Cortes and Vapnik, 1995) solve the following problem:

$$\min_{w,b,\xi} \frac{1}{2}w^T w + C\sum_{i=1}^{l}\xi_i$$

$$\text{subject to } y_i(w^T\phi(x_i)+b) \geq 1-\xi_i$$

$$\xi_i \geq 0, i = 1,...,l$$

Vectors x_i are mapped in a space with more dimensions than the initial one, so the SVM find an iperplane of separation with the largest margin possible in this new space.

C (>0) is the penalty parameter of the error. Furthermore, $K(x_i, x_j) = \varphi(x_i)^T\varphi(x_j)$ is called the kernel function and we used as kernel fuction the radial basis fuction (RBF):

$$K(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|^2), \quad \gamma > 0$$

In order to find the best network we adepte the procedure by Hsu Chih-Wei, Chih-Chung Chang and Chih-Jen Lin (creators of LIBSVM);

Since the two parameters (C and Gamma) cannot be estimated in advance we carried out 110 training experiments, letting C and Gamma vary as C = 2^{-5}, 2^{-3}, ... 2^{15}, Gamma = 2^{-15}, 2^{-13}… 2^3 (a factor of 4 between a value and the next). Training was performer using cross validation and by dividing the dataset in 5 shares.

Given the weight, it is virtually impossible to run these processes in series on a desktop, and it was therefore decided to use the GRID technology, using the PONs (SCoPE, COMETA and Cybersar) GRID infrastructure. Each Experiment used 110 worker nodes of the GRIDs at the same time.

Sample	Parameters	BoK	Algorithm	e_{tot}	c_{tot}
Experiment (1)	SDSS photometric parameters + photo redshift	BPT plot +Kewley's line	SVM / MLP	~74% / ~76%	~55% / ~54%
Experiment (2)	SDSS photometric parameters + photo redshift	BPT plot+Kewley's line	SVM / MLP	e_{typ1}~82% e_{typ2}~86% / e_{typ2}~98% e_{typ1}~95%	~98% / ~100%
Experiment (3)	SDSS photometric parameters + photo redshift	BPT plot+Heckman's+Kewley's lines	SVM / MLP	~78% / ~80%	~89% / ~92%