# Photometric classification of emission line galaxies with machine-learning methods

Stefano Cavuoti,[1,2]★ Massimo Brescia,[1] Raffaele D'Abrusco,[3] Giuseppe Longo[2,4] and Maurizio Paolillo[2]

[1]*INAF-Astronomical Observatory of Capodimonte, via Moiariello 16, I-80131 Napoli, Italy*
[2]*Department of Physical Sciences, University Federico II, via Cinthia 6, I-80126 Napoli, Italy*
[3]*Harvard–Smithsonian Center for Astrophysics, Cambridge, MA 02138, USA*
[4]*Visiting Associate, California Institute of Technology, Pasadena, CA 91125, USA*

## ABSTRACT

In this paper, we discuss an application of machine-learning-based methods to the identification of candidate active galactic nucleus (AGN) from optical survey data and to the automatic classification of AGNs in broad classes. We applied four different machine-learning algorithms, namely the Multi Layer Perceptron, trained, respectively, with the Conjugate Gradient, the Scaled Conjugate Gradient, the Quasi Newton learning rules and the Support Vector Machines, to tackle the problem of the classification of emission line galaxies in different classes, mainly AGNs versus non-AGNs, obtained using optical photometry in place of the diagnostics based on line intensity ratios which are classically used in the literature. Using the same photometric features, we discuss also the behaviour of the classifiers on finer AGN classification tasks, namely Seyfert I versus Seyfert II, and Seyfert versus LINER. Furthermore, we describe the algorithms employed, the samples of spectroscopically classified galaxies used to train the algorithms, the procedure followed to select the photometric parameters and the performances of our methods in terms of multiple statistical indicators. The results of the experiments show that the application of self-adaptive data mining algorithms trained on spectroscopic data sets and applied to carefully chosen photometric parameters represents a viable alternative to the classical methods that employ time-consuming spectroscopic observations.

**Key words:** methods: data analysis – catalogues – surveys – galaxies: active – galaxies: Seyfert.

## 1 INTRODUCTION

Active galactic nuclei (AGNs) and their high-redshift counterpart, quasars, are the most luminous long-lived discrete objects in the Universe and are therefore crucial to address a variety of astrophysical and cosmological problems. Individual multiwavelength studies allow to investigate the physical conditions in the proximity of the central power source, a supermassive black hole with a surrounding accretion disc. The study of well-defined samples of AGNs in various environments, both in the local universe and at high redshift, is needed to constrain the various mechanisms invoked to explain galaxy assembly and early evolution (Mahajan, Haines & Raychaudhury 2010). It is also needed to explore the role of the environment in triggering or inhibiting nuclear activity (Kauffmann et al. 2004; Popesso & Biviano 2006).

In spite of the fact that the unified model (Antonucci 1993) may provide a unique physical explanation for the central engine and the surrounding regions, the observable phenomenology of AGNs (and quasars) is quite complex and encompasses a variety of objects (Seyfert galaxies, LINER, quasars, blazars, etc.), which, for the large differences existing in their observed properties, have for a long time been considered to be different and independent species (Antonucci 1993).

This phenomenological complexity is also the reason why there cannot be a unique method equally effective in identifying all AGN phenomenologies (cf. Messias et al. 2010) in every redshift range. While it is clear that the most effective and the most used method relies on X-ray emission properties (cf. Alexander et al. 2002), other types of indicators can be effectively used such as mid-infrared fluxes (cf. Donley et al. 2007), colour–colour diagrams (cf. Hatziminaoglou et al. 2005) and even radio data both through peak emission and through unresolved radio emission (Seymour 2007). All these methods have their share of pro and cons and are more or less biased against the detection of specific subtypes.

The techniques developed to classify galaxies based on the presence and intensity of spectral emission features have received special attention in the astronomical literature. Emission lines are

★ E-mail: stefano.cavuoti@gmail.com

visible in the spectra of a large fraction of galaxies of different classes. The relative strengths of multiple emission lines have been successfully used to infer the class of galaxies observed spectroscopically (see, for example, the ground-breaking work discussed by Veilleux & Osterbrock 1987). Among emission line galaxies, Seyferts, LINERs and starburst galaxies can be distinguished based on their characteristic positions in the diagrams generated by the ratios of the equivalent width of the [O III](5007), H$\beta$ lines on the *y*-axis and [N II](6584), H$\alpha$ lines on the *x*-axis (Veilleux & Osterbrock 1987). In this diagram, starburst galaxies are located in the lower left-hand region, narrow-line Seyferts are located in the upper-right corner and LINERs tend to occupy the lower right-hand region. Different parametric models of the lines delimiting these different regions have been proposed based on the growing availability of large samples of galaxies with spectra (Kewley et al. 2001; Lamareille 2010). Details on the most recent parametrization of the line ratio diagnostic diagram are given in Section 2. While such methods based on spectroscopic data provide efficient and reliable classification of line-emitting galaxies, they are applicable only to galaxies with measured spectra, which usually represent a small fraction of the total number of galaxies observed in the modern mixed (photometric and spectroscopic) digital optical surveys (e.g. the Sloan Digital Sky Survey – SDSS; see York et al. 2000).

In this paper, we investigate the possibility to: (i) identify candidate AGNs and, (ii) classify them in broad classes using optical photometric parameters only, by means of supervised classification techniques trained on samples of spectroscopically classified galaxies. This study was motivated by the growing number of planned and ongoing optical surveys covering, with high accuracy and depth, large portions of the sky, such as KIDS (de Jong et al. 2013), DES (Annis 2013), PANSTARRS (Tonry et al. 2012), etc., where the possibility to identify reliable AGN samples from optical data alone, would be very helpful in selecting well-characterized statistical samples and to identify candidates for subsequent spectroscopic validation and other follow-up studies.

The new digital surveys produce complex data with many tens or hundreds of parameters measured by automatic methods and carry much more information than in the past. This wealth of accurate data has opened the path to the use of data mining methods (typical of machine learning) in place of the usual statistical tools to perform all sorts of classification, regression and clustering.

An interesting attempt to tackle this problem was performed by Suchkov, Hanisch & Margon (2005) who tried to reproduce the SDSS classification using only optical colours and reached the conclusion that *SDSS colours feature prominently in the algorithm used to select AGN candidates for subsequent SDSS spectroscopy* (Suchkov et al. 2005).

The work described here was performed using supervised machine-learning methods offered to the community through the DAta Mining & Exploration Web Application REsource (DAMEWARE[1]). For those who are not familiar with the topic, we shall just remind that supervised methods learn how to perform a given operation (for instance how to disentangle normal from active galaxies) using a set of well-known examples (also known as a priori knowledge or *knowledge base*). The methods we used in this paper are, respectively, Support Vector Machines (SVM; Chang & Lin 2001) and the Multi Layer Perceptron (MLP) with different types of training rules: the Conjugate gradient (CG; Golub & Ye 1999), the Scaled Conjugate Gradient (SCG; Watrous 1987) and the

Quasi Newton Algorithm (MLPQNA; Brescia et al. 2012) and are shortly summarized in Section 3.

The data set used for the experiments was obtained by joining three catalogues of objects (within the redshift range $0.02 < z < 0.3$), respectively, from Sorrentino, Radovich & Rifatto (2006), Kauffmann et al. (2003) and D'Abrusco et al. (2007), as described in Section 2. The data set contains the photometric parameters (hereinafter named as input features), as well as flags describing the nature of the objects (in machine-learning methods, this flag is also called target) to be used only in the training and test steps of any experiment. These flags constitute the Knowledge Base (or KB).

We performed three types of experiments. First of all, the detection of candidate AGNs (AGN versus non-AGN), which is the main experiment; then we tried to classify Seyfert I against Seyfert II type galaxies; finally, we also tested the possibility to distinguish Seyfert galaxies against LINERs. The data sets used in the experiments and the spectroscopic parametrization used to train the classification methods are described in Section 2, while the detailed description of the different classes of algorithms used is given in Section 3. The results of the experiments and the performances of the four techniques to each different class of experiment are described in Section 4 and we discuss our findings in Section 5. Finally, we draw our conclusions in Section 6.

## 2 THE KB AND THE DATA

Supervised methods learn how to reproduce the desired knowledge using the already mentioned KB, i.e. a collection of *examples*, that are patterns for which the right classification (target) is already known by independent means. It goes without saying that a biased, incomplete or poor KB will affect the classification efficiency. In other words, since one of the main drawbacks of the machine-learning methods is the difficulty in extrapolating to regions of the input parameter space (PS) that are not well sampled by the training data, the KB needs to cover in a homogeneous way the whole PS with a local density which depends on the complexity of the knowledge to be reproduced.

For the reasons outlined in the previous paragraph, it is quite evident that, in the case of AGNs, the construction of a complete and unbiased KB is almost an impossible dream unless very conservative choices, such as those adopted in this paper, are made.

Our KB was obtained by merging two different samples (respectively, Sorrentino et al. 2006 and Kauffmann et al. 2003) of objects for which a classification based on spectroscopy was available.

Both samples were drawn from the SDSS Data Release 4 (DR4) *PhotoSpecAll* table which contains all objects for which both photometric and spectroscopic observations are available.

Catalogue by Sorrentino et al. (2006). This catalogue contains objects in the redshift range $0.05 < z < 0.095$. It provides a classification as type 1 (Seyfert I and LINER I), type 2 (Seyfert II and LINER II) and non-AGN for 24 293 objects. The data were extracted from the SDSS DR4 (Adelman-McCarthy et al. 2006), and the selection was performed using the traditional approach based on the equivalent width of specific emission lines. In particular, objects classification was originally performed by Sorrentino et al., which are assumed to be bona fide AGN sources that lay above one of the so-called Kewley's lines (Kewley et al. 2001):

$$\log \frac{[\text{O III}]\lambda 5007}{\text{H}\beta} = \frac{0.61}{\log \frac{[\text{N II}]\lambda 6583}{\text{H}\alpha} - 0.47} + 1.19 \qquad (1)$$

---

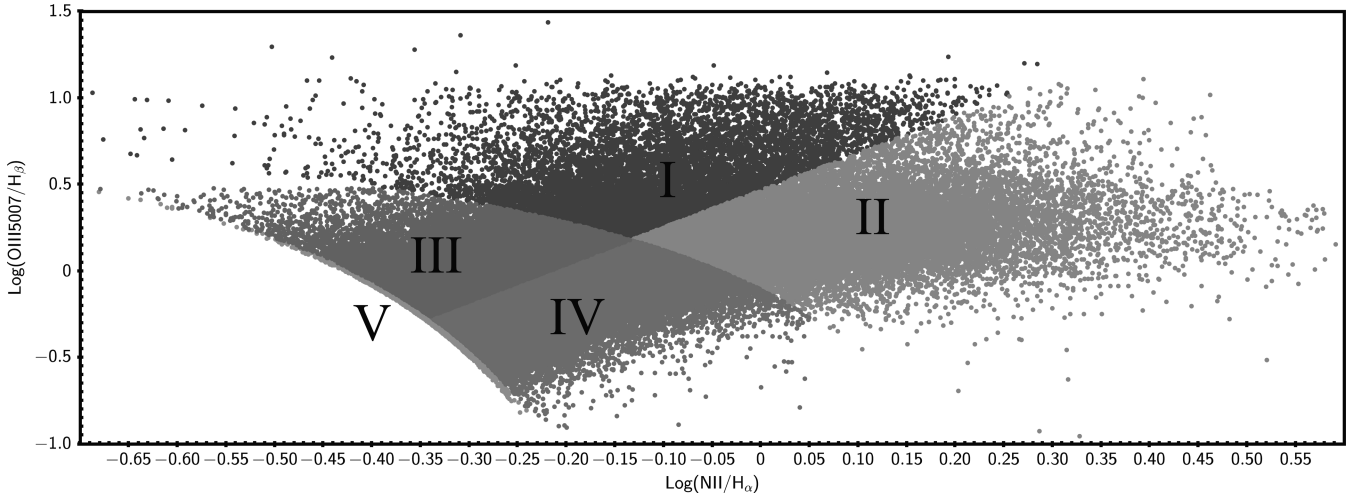[1] http://dame.dsf.unina.it/beta_info.html

**Figure 1.** A representation of the Kauffmann catalogue in the BPT diagram. The blue (I) region is populated by Seyferts; green region (II) by LINERs; the violet (III) by a mixture of Seyferts and non-AGN; the grey region (IV) is populated by a mixture of non-AGN and LINERs; finally, light blue region (V) at the lower-left boundary is populated by non-AGN.

$$\log \frac{[\text{O\,III}]\lambda5007}{\text{H}\beta} = \frac{0.72}{\log \frac{[\text{S\,II}]\lambda\lambda6717,6731}{\text{H}\alpha} - 0.32} + 1.30 \qquad (2)$$

$$\log \frac{[\text{O\,III}]\lambda5007}{\text{H}\beta} = \frac{0.73}{\log \frac{[\text{O\,I}]\lambda6300}{\text{H}\alpha} - 0.59} + 1.33 \qquad (3)$$

Furthermore, AGNs were classified as Seyfert I if:

$$\text{FWHM}(\text{H}\alpha) > 1.5\text{FWHM}([\text{O\,III}]\lambda5007) \qquad (4)$$

or

$$\text{FWHM}(\text{H}\alpha) > 1200 \,\text{km s}^{-1} \qquad (5)$$

and

$$\text{FWHM}([\text{O\,III}]\lambda5007) < 800 \,\text{km s}^{-1}. \qquad (6)$$

All the other AGNs were classified as Seyfert II. The final catalogue comprises of 22 464 objects recognized as non-AGN, 725 Seyfert I and 1105 Seyfert II (see their figs 2 and 3).

Catalogue by Kauffmann et al. (2003). This catalogue[2] contains spectra lines and ratio for 88 178 galaxies ($0.02 < z < 0.3$). Since this is a purely spectroscopic catalogue, in order to divide objects in different classes, we followed Kauffmann et al. (2003) prescriptions defining a region populated by AGNs above the Kewley line, (equation 1 Kewley et al. 2001), and a second region populated by objects, which are likely not to be AGN below the Kauffmann line (Kauffmann et al. 2003; Kewley et al. 2006):

$$\log \frac{[\text{O\,III}]\lambda5007}{\text{H}\beta} = \frac{0.61}{\log \frac{[\text{N\,II}]\lambda6583}{\text{H}\alpha} - 0.05} + 1.3. \qquad (7)$$

The intermediate region is heavily contaminated by non-AGN and, in what follows, we shall refer to it as the *mixed* zone.

Finally, as in the previous case, we used the following Heckman's line (Heckman 1980; Kewley et al. 2006):

$$\frac{[\text{O\,III}]\lambda5007}{\text{H}\beta} = 2.1445 \frac{[\text{N\,II}]\lambda6583}{\text{H}\alpha} + 0.465 \qquad (8)$$

to further divide the sample in Seyferts and LINERS.

[2] http://www.mpa-garching.mpg.de/SDSS/DR4/

The resulting five areas in the plane defined by the equivalent widths line ratios (Fig. 1) are populated by objects from the Kauffmann sample classified according to the Baldwin, Phillips and Terlevich (BPT; Baldwin, Phillips & Terlevich 1981) diagram diagnostics. In the diagram, it is evident that this catalogue contains a few objects below the Kauffmann line (surely non-AGN), although other non-AGN objects are present in the mixed zone and in the catalogue by Sorrentino et al. (2006).

Catalogue by D'Abrusco et al. (2007). This catalogue[3] contains photometric redshifts for all SDSS DR4 with $z < 0.4$, matching the following selection criteria: dereddened magnitude in the $r$ band, $r < 21$; mode = 1 which corresponds to primary objects only in the case of deblended sources.

The first two catalogues were merged together and, in the case of overlapping entries, we retained the type 1 and type 2 information from Sorrentino et al. (2006) and all other types from Kauffmann et al. (2003). The resulting merged catalogue included a total of 108 162 objects. The two catalogues have ∼2000 common objects. In principle, it could be possible to perform the same classification performed by Sorrentino et al. (2006) on the Kauffmann et al. (2003) catalogue. The choice has been driven by considering their divergent and specific goal: Sorrentino et al. (2006) tried to investigate the AGN environment by following a very conservative approach; on the other side, the prescriptions of Kauffmann et al. (2003) and Kewley et al. (2006) are more general purpose. It was finally cross-matched with the third catalogue, containing photometric redshifts (photo-$z$) provided by D'Abrusco et al. (2007), having a very low standard deviation of residual error ($\sigma \simeq 0.02$), which further reduced the number of objects to 100.069. The need for photo-$z$ was dictated by our goal to identify potential AGN objects using only the available photometric information (a choice which ruled out the use of more accurate spectroscopic redshifts). For these objects, we extracted the following parameters from the SDSS archive and for each band (see SDSS website[4] for details):

(i) *fiber Mag*, flux in 3 arcsec diameter fibre radius;
(ii) *petro Mag*, Petrosian flux;

[3] http://dame.dsf.unina.it/catalogues.php
[4] http://www.sdss.org/

**Table 1.** The final data set (KB) composition. Empty fields stand for the unused typology. The separation between class 0 and class 1 are referred to the target vector (used during training).

| CLASS | Catalogue | Exp. AGN versus non-AGN | Exp. Seyfert I versus Seyfert II | Exp. Seyfert versus LINER |
|---|---|---|---|---|
| Non-AGN | All | Class 0 | – | – |
| Type 1 | Sorrentino | Class 1 | Class 1 | – |
| Type 2 | Sorrentino | Class 1 | Class 0 | – |
| Mix-LINER | Kauffmann | Class 0 | – | – |
| Mix-Seyfert | Kauffmann | Class 0 | – | – |
| Pure-LINER | Kauffmann | Class 1 | – | Class 0 |
| Pure-Seyfert | Kauffmann | Class 1 | – | Class 1 |
| Mix-LINER-Type 1 | Overlap | Class 0 | – | – |
| Mix-Seyfert-Type 1 | Overlap | Class 0 | – | – |
| Pure-LINER-Type 1 | Overlap | Class 1 | Class 1 | Class 0 |
| Pure Seyfert-Type 1 | Overlap | Class 1 | Class 1 | Class 1 |
| Mix-LINER-Type 2 | Overlap | Class 0 | – | – |
| Mix-Seyfert-Type 2 | Overlap | Class 0 | – | – |
| Pure-LINER-Type 2 | Overlap | Class 1 | Class 0 | Class 0 |
| Pure-Seyfert-Type 2 | Overlap | Class 1 | Class 0 | Class 1 |
| SIZE: 24293 | Sorrentino | 84 885 | 1570 | 30 380 |
| SIZE: 88178 | Kauffmann | – | – | – |

(iii) *petroR*50, radius containing 50 per cent of Petrosian flux;
(iv) *petroR*90, radius containing 90 per cent of Petrosian flux;
(v) *dered*, dereddened magnitude, corrected for extinction.

Therefore, the number of initial parameters is 26 (i.e. the five SDSS parameters listed above for each of the five SDSS *ugriz* bands plus the photometric redshift). We also used derived parameters such as colours and concentration index. Finally, all objects with undefined values for some of their parameters (named also as *Not a Number* or NaN) were removed and this reduced the final number of objects to 84 885. In this case, with the term *undefined values* we mean undefined numerical values underlying either non-detection or contaminated measurements. This last step is crucial in machine-learning methods since the presence of such unknown data might affect their generalization capabilities (Marlin 2008).

This final catalogue, summarized in Table 1, was then used to create three different data sets to be used for the three distinct classification experiments described in Section 4. Namely:

(i) KB data set for the AGN versus non-AGN experiment: the whole (Kauffmann + Sorrentino) catalogue;

(ii) KB data set for the Seyfert I versus Seyfert II experiment: just the pure AGN objects belonging to the data set of Sorrentino et al. (2006) resulting into 1570 objects;

(iii) KB data set for the Seyferts versus LINERs experiment: pure AGN objects, belonging to the catalogue of Kauffmann et al. (2003), divided into LINERs and Seyferts, obtaining 30 380 objects.

## 3 THE METHODS

DAMEWARE (cf. Brescia et al. 2010) is among the main products made available through the DAME (Data Mining and Exploration) Program Collaboration. It provides a web application, able to configure and execute data mining experiments through machine-learning models on a distributed computing infrastructure. More recently, some of these machine-learning algorithms were offered in a parallel version which exploits the computing possibilities offered by Graphical Processing Unit technology (Cavuoti et al. 2013).

From the models available in DAMEWARE, we selected four supervised classifiers, the SVM and three variants of the MLP, a stan-

dard neural network (NN) trained by different types of self-adaptive learning rules, respectively, the CG, the SCG and the MLPQNA.

SVM (Chang & Lin 2001): SVM are supervised learning models with associated learning algorithms that analyse data and recognize patterns which are mostly used for classification and regression analysis. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier. Given a set of training examples, each marked as belonging to one of the two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and assigned to a category depending on which side of the gap they fall on. In addition to performing linear classification, SVMs can efficiently perform non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. Among different types of such kernel function, we used the radial basis function type (Chang & Lin 2001).

The SVM training experiments over large data sets have huge computational cost (about one week per experiment on a single CPU for a data sample of about 80 000 input patterns); thus, in order to be able to perform the hundreds of experiments described in what follows (see Section 4), it was needed to exploit the SCoPE[5] GRID infrastructure resources (Brescia et al. 2009).

MLP (Bishop 1995): NNs have long been known to be excellent tools for interpolating data and for extracting patterns and trends and since a few years they have also carved their way into the astronomical community for a variety of applications (see the reviews by Tagliaferri et al. 2003a,b and references therein), ranging from star–galaxy separation (Donalek 2006), spectral classification (Winter, Jeffery & Drilling 2004) and photometric redshifts evaluation (Cavuoti et al. 2012; Brescia et al. 2013). In practice, an NN is a tool which takes a set of input values (input neurons), applies a non-linear (and unknown) transformation and returns an output. The optimization of the output is performed by using a set of examples

[5] http://www.scope.unina.it/C19/astrophysics-gridcomputing

for which the output value (target) is known a priori. Performances of an MLP are greatly affected by the choice of the learning rule, i.e. by the mathematical expression used for the optimization of its internal weights. In this paper, we tested three different rules, namely the CG, the SCG and the MLPQNA. In essence, the learning process of an MLP consists of two phases through the different layers of the network: a forward pass and a backward pass. In the forward pass, an input vector is applied to the input nodes of the network, and its effect propagates through the network layer by layer. Finally, a set of outputs is produced as the actual response of the network. During the backward pass, on the other hand, the weights are all adjusted in accordance with the error-correction rule. The training of NNs, such as MLP, implies to find the more efficient among a population of NNs differing in the hyperparameters controlling the learning of the network, in the number of hidden nodes, etc. The most important hyperparameter (usually called $\alpha$) is related to the weights of the network and allows to estimate the dependence of the training performance on the different inputs and the selection of the parameters for a given task. In fact, a larger value of $\alpha$ implies a less meaningful corresponding weight (Bishop 1995). The three variants of learning rules discussed here differ basically in the way to calculate the $\alpha$ parameter.

All these algorithms found the minimum of a square error function, but the computational cost of each step is high, because in order to determine the values of $\alpha$, we have to refer to the *Hessian matrix* **H** of the error, which is highly expensive in terms of calculations. But fortunately, the coefficients, such as the parameter $\alpha$, can be obtained from analytical expressions that do not use the Hessian matrix explicitly. The method of CG reduces the number of steps to minimize the error up to a maximum of $|w|$ (where $|w|$ is the cardinality of network weights), because there could be almost $|w|$ conjugate directions in a $|w|$-dimensional space (Golub & Ye 1999). The SCG method differs from the CG by imposing that the Hessian matrix **H** is always positive (Nocedal & Wright 1999). This can be done by adding to **H** a multiple of identity matrix $\lambda I$, where $I$ is the identity matrix and $\lambda > 0$ is a scaling coefficient (Watrous 1987). Finally, the MLPQNA does not calculate the **H** matrix, but an approximation in a series of steps. A famous implementation of the QNA, which offers good performance even for non-smooth optimizations, is known as BFGS, by the names of its inventors (Broyden 1970; Fletcher 1970; Goldfarb 1970; Shanno 1970), and it was our choice. This approach generates a sequence of matrices *G* which are subsequently more and more accurate approximations of the Hessian matrix by using only information related to the first derivative of error function (Brescia et al. 2012).

In what follows, we outline the standard data mining procedure which was adopted in all the AGN classification experiments performed with the machine-learning models described above.

### 3.1 Feature extraction

The first step is the pruning of the input parameters. Most machine-learning methods are in fact quite demanding in terms of computing time, which may scale badly with the number of input parameters (features). It is therefore necessary to optimize the number of input features by performing what is usually called the *feature selection or pruning* phase aimed at identifying the subset of features carrying the highest amount of information for a specific task.

In order to perform the pruning of input features, the initial 26 features have been organized by replacing the five magnitudes *dered* for each band with the corresponding colours plus the *r dered* magnitude as reference, due to their capability to improve the performance,

as revealed after some preliminary experiments. The improvement carried by colours can be easily understood by noticing that even though colours are derived as a subtraction of magnitudes, the content of information is quite different, since an ordering relationship is implicitly assumed, thus increasing the amount of information in the final output (gradients instead of fluxes). The additional reference magnitude instead removes the degeneracy in the luminosity class for a specific galaxy type (Brescia et al. 2013).

Then a *leave-one-out* cyclic method has been used to test the contribution of each single feature to the classification training performance and to remove each time the worst resulting. This cyclic procedure is stopped when the performance does not increase by removing any further feature.

The leave-one-out procedure was performed according to the following top-down strategy:

(i) perform one experiment with all the features and store the performance;

(ii) perform $N$ experiments, where $N$ is the whole number of features in the data set, by removing each time one of the features;

(iii) find the set of features achieving the best performance;

(iv) if the achieved performance is better or equal than the previous one, remove the feature from the set, store the result and go back to point (ii), otherwise stop the procedure.

At the end, the feature extraction phase produced the following subset of seven selected input features, candidates to perform the final classification experiments.

(i) the 4 SDSS colours $(u - g), (g - r), (r - i), (i - z)$, dereddened for galactic absorption;

(ii) the dereddened magnitude in the $r$ band;

(iii) *fiberMag_r* the fibre magnitude in the $r$ band;

(iv) the photometric redshift derived from D'Abrusco et al. (2007).

All these features have been used in all the experiments in order to maintain the coherence along the overall classification process.

### 3.2 Model architecture selection

The second step consists in identifying for each model, via a trial-and-error procedure, the best architecture, which, for instance, in the case of MLP would mean to find the optimal number of neurons in the hidden layer, the optimal learning function, etc. Since there is no way to define it a priori, it is necessary to perform many experiments changing every time the parameters defining the model.

In each experiment, the KB is randomly split into two parts, namely the training set (70 per cent of the data set) to be used by the model to learn the classification rule and the test set (the remaining 30 per cent), used exclusively to evaluate the results. Due to the supervised nature of the classification task, the system performance can be measured by means of a test set during the testing procedure, in which unseen data are given to the system to be labelled. The overall performance thus integrates information about the classification accuracy (i.e. in terms of output correctness). Moreover, the results obtained from the unseen data are also important to evaluate the learning robustness, i.e. the generalization capability of the network in presence of data samples never used during the training phase.

Furthermore, it is important to stress that, in order to ensure a proper coverage of the KB in the PS, the data objects were indeed divided among the training and blind test data sets by random extraction, and this process minimizes the possible biases induced by

statistical fluctuations in the coverage of the PS, namely small differences in the class distribution of training and test samples used in the experiments.

### 3.3 Evaluation of performances

Performances were evaluated on the test set using a standard set of statistical indicators defined in this section. We wish to stress that the test were performed by submitting to any given model the photometric data alone and then by comparing the predicted value with the target. For a given confusion matrix:

$$
\begin{array}{c|cc}
 & \multicolumn{2}{c}{\text{OUTPUT}} \\
 & \text{Class A} & \text{Class B} \\
\hline
\text{TARGET} \quad \text{Class A} & N_{AA} & N_{AB} \\
\text{Class B} & N_{BA} & N_{BB},
\end{array} \tag{9}
$$

we can define the following statistical quantities.

(i) Total efficiency: te. Defined as the ratio between the number of correctly classified objects and the total number of objects in the data set. In our confusion matrix example, it would be

$$
te = \frac{N_{AA} + N_{BB}}{N_{AA} + N_{AB} + N_{BA} + N_{BB}}. \tag{10}
$$

(ii) Purity of a class: pcN. Defined as the ratio between the number of correctly classified objects of a class and the number of objects classified in that class, also known as efficiency of a class. In our confusion matrix example, it would be

$$
pcA = \frac{N_{AA}}{N_{AA} + N_{BA}}, \tag{11}
$$

$$
pcB = \frac{N_{BB}}{N_{AB} + N_{BB}}. \tag{12}
$$

(iii) Completeness of a class: cmpN. Defined as the ratio between the number of correctly classified objects in that class and the total number of objects of that class in the data set. In our confusion matrix example, it would be

$$
cmpA = \frac{N_{AA}}{N_{AA} + N_{AB}}, \tag{13}
$$

$$
cmpB = \frac{N_{BB}}{N_{BA} + N_{BB}}. \tag{14}
$$

(iv) Contamination of a class: cntN. It is the dual of the purity, namely it is the ratio of misclassified object in a class and the number of objects classified in that class, in our confusion matrix example will be

$$
cntA = 1 - pcA = \frac{N_{BA}}{N_{AA} + N_{BA}}, \tag{15}
$$

$$
cntB = 1 - pcB = \frac{N_{AB}}{N_{AB} + N_{BB}}. \tag{16}
$$

## 4 EXPERIMENTS AND RESULTS

We performed three different kinds of experiments: (i) AGN detection; (ii) Seyfert I versus Seyfert II classification and (iii) Seyferts versus LINERs classification. In all cases, the experiments were approached with two kinds of machine-learning models, respectively, SVM and MLP, the latter in three different versions, by changing the internal learning rule (i.e. CG, SCG and QNA) as described in Section 3.

### 4.1 AGN classification

Concerning the classification of AGN against non-AGN, the MLP models were trained using a target vector whose values were set to 1 for each object above the Kewley line (i.e. pure AGN) and to 0 for object below it (which therefore includes the mixed zone objects which are non-AGNs). The KB included 84 885 objects after the removal of the patterns affected by NaNs. According to the mentioned strategy, the training set (70 per cent of the whole data set) contained 59 419 patterns while the test set (30 per cent of the data set) contained 25 466 patterns.

The MLP output may be interpreted as the probability for a given object to belong to a specific class and a threshold needs to be assumed in order to classify the objects. With the standard choice of such threshold to be 0.5, for instance, an object above the threshold is considered to belong to the class of AGNs. Such threshold represents the median point of the probability to assign the MLP output to a class or another in a two-class classification problem. As can be seen from Table 2, which summarizes the results from all the experiments, the best result was obtained by the MLP with the Quasi Newton learning rule.

In the best experiment, the SVM reached a comparable result, 75.8 per cent, obtained with $C = 32\,768$ and $\gamma = 0.001\,953\,125$, where $C$ is a penalty parameter and $\gamma$ the internal parameter of the radial basis function kernel (Chang & Lin 2001). Table 2 reports the complete results by using the three MLP learning rules and the SVM.

### 4.2 Seyfert I versus Seyfert II classification

In the classification between type I and type II Seyfert objects, the ML models were fed using a target vector whose values were set to 1 for objects classified as Seyfert I in the catalogue by Sorrentino et al. (2006) and to 0 if classified as Seyfert II resulting in 1830 objects and 1570 after the usual removal of the patterns affected by NaN values. So, the training set contained 1256 patterns while the test set 314 patterns.

**Table 2.** AGN versus non-AGN: the first column is the model used, while the others give, respectively, the total efficiency, the AGN completeness, the non-AGN completeness, the AGN purity and the AGN contamination. These percentages are calculated by considering only the results on the 25 466 objects of test set (i.e. not including training set results).

| Model | te | cmpAGN | cmpMIX | pcAGN | cntAGN |
|-------|-----|--------|--------|-------|--------|
| CG | 75.5 per cent | 55.6 per cent | 86.3 per cent | 68.5 per cent | 31.5 per cent |
| SCG | 75.7 per cent | 55.1 per cent | 86.2 per cent | 68.4 per cent | 31.6 per cent |
| QNA | 76.5 per cent | 58.6 per cent | 86.5 per cent | 70.8 per cent | 29.2 per cent |
| SVM | 75.8 per cent | 55.4 per cent | 86.2 per cent | 70.6 per cent | 29.4 per cent |

**Table 3.** AGN versus non-AGN: comparison among the seven features of the test set in terms of their amount of information given to the classification performance. The first column reports the excluded features from the training set for each pruning experiment. The others are, respectively, total efficiency, completeness, purity and contamination. The values are expressed in terms of percentage variations with respect to the best values (QNA) reported in Table 2.

| Excluded features | te | cmp | pc | cnt |
|---|---|---|---|---|
| photo-$z$ | −0.4 per cent | −0.8 per cent | −0.6 per cent | +0.6 per cent |
| *fiberMag_r* | −0.9 per cent | −2.8 per cent | −1.7 per cent | +1.7 per cent |
| *dered_r* | −0.6 per cent | −2.5 per cent | −0.3 per cent | +0.3 per cent |
| all except colours | −0.9 per cent | −1.6 per cent | −3.7 per cent | +3.7 per cent |

In this case, the main parameter of interest that quantifies the ability to distinguish the two classes is the efficiency; the MLPQNA model produced a total efficiency of 72 per cent, while using the SVM, the best result produced a total efficiency equal to 81.5 per cent. As can be seen, the results are promising even more if we take into account the small number of patterns used for the training.

### 4.3 Classification of Seyferts versus LINERs

Concerning the last experiment, namely the classification between Seyfert and LINER objects, the ML models were fed using a target vector with values labelled as 1 for objects lying below the Heckman line and 0 for objects above the line. This resulted in a total of 30 380 objects after the removal of the patterns affected by NaN presence. The training set (70 per cent of the whole data set) contained 21 266 patterns and the test set 9114 patterns.

The MLPQNA model produced a total efficiency of 73.8 per cent, while using SVM, we reached the best total efficiency equal to 78.18 per cent.

## 5 DISCUSSION OF THE EXPERIMENTS

In general terms, in the main experiment (i.e. the classification AGN versus non-AGN), the MLP with QNA learning rule performs better than all other methods, both in terms of performance and robustness. This is not completely a surprise since already in other cases, the MLPQNA has been proven (Brescia et al. 2012) to be quite effective in optimizing the poor information introduced by a small or incomplete KB due to its fine approximation of the Hessian of the training error (Broyden 1970; Fletcher 1970; Goldfarb 1970; Shanno 1970). We obtain a good overall efficiency, of about 75 per cent with a good purity (70 per cent) while all methods performed badly in terms of completeness reaching about 58 per cent even though it must be stressed that if a high level of purity is needed for specific applications, MLPQNA can be fine-tuned to do it by varying the threshold at which an object is recognized as AGN, but this can be done at the price of a loss in completeness. The low completeness may be partly explained by the ambiguities introduced by template patterns in the mixed zone. We therefore investigated the possibility to increase the purity at a relative price of completeness by changing the threshold level. The optimal value has been obtained for the threshold 0.87 which leads to a purity of 88 per cent and a level of completeness of 9 per cent. It goes without saying that the balancing between purity and completeness can be performed according to the need from the specific application.

By considering the data set which gives the best results, obtained with the MLPQNA model, we performed a series of experiments to evaluate the contribution of each feature of training objects to the test performance, in terms of information given to the classification during training. This set of tests has been done by alternately excluding some of the features for all training objects. The resulting variation percentages for all used statistical indicators are shown in Table 3. We emphasize that in our case photometric redshifts are crucial to reproduce the same cut at spectroscopical redshift <0.3, imposed by the original KB (Kauffmann et al. 2003). This is a typical requirement of empirical methods in order to maintain the coherence between trained and new data samples in terms of PS.

Concerning the analysis of the contribution to the classification performance of the photometric features, composing the training and test patterns, the series of tests, reported in Table 3, have shown a significant valence of colours and reference magnitudes (mainly *fiberMag* but also *dered* in the *r* band), followed by an important contribution of photometric redshift. Although not surprising for colours, due to their objective quantity of correlated information carried, it resulted quite interesting that without information given by photo-$z$ and reference magnitudes, the classification capability underwent a significant decrease.

By considering the subset of non-AGN objects within the class including both mixed and non-AGN objects, its percentage of false positives (i.e. those misclassified as AGN) is about 1 per cent. Moreover, the percentage of objects, spectroscopically known as non-AGN, which become false positive, is also about 1 per cent. The contamination due to galaxies is very small, and this must be considered very encouraging since the ambiguities in the KB, introduced by unrecognized AGN in the mixed zone, can only lower this already very small percentage.

Concerning the experiment related to the classification of objects in Seyfert I versus Seyfert II (hereinafter experiment 2), the level of performance can be easily understood in terms of the small dimension of the training data set, since in general machine-learning techniques are quite sensitive to the incompleteness of the KB.

About the classification Seyfert versus LINER (hereinafter experiment 3), the contamination in the lower region near to the Heckman line *confuses* the machine-learning techniques, leading, in turn, to reduced performances of the photometric classification. This is also partially true in the first experiment (AGN versus non-AGN), where a contamination is also present in the mixed zone, i.e. between Kauffmann and Kewley lines.

Hence, a clear result of our experiments is that an unambiguous KB is required to successfully train and apply any classification method. This can be brought back to the fact that Seyfert I and Seyfert II, from the optical photometry point of view, show substantially different behaviour, while the difference between Seyfert and LINER is somehow more vague; this situation is even worse in the AGN versus non-AGN experiment where the whole mixed zone *confuses* the network. This is also evident in the spectroscopic PS

where the so-called *seagull wings* move far away from the Kewley line. By considering the Seyfert I and II alone, it results evident a quite sharp spectroscopical separation (Sorrentino et al. 2006).

## 6 CONCLUSIONS

The production of large and accurate AGN catalogues is an important topic that will become crucial with the advent of the future photometric only digital surveys that will map large fractions of the sky to unprecedented depth in the different wavelengths.

We have applied four distinct classification methods, based on self-adaptive classification techniques, to the problem of the classification of emission line galaxies using only optical photometric parameters. The methods have been applied to three classification problems, specifically the separation of AGNs from non-AGNs, Seyfert I from Seyfert II and the classification of Seyfert from LINERs. In terms of classification efficiency, the results indicate that our methods perform fairly ($\sim$76.5 per cent) when applied to the problem of the classification of AGNs versus non-AGNs, while the performances in the more fine classification of Seyfert versus LINERs are $\sim$78 per cent and $\sim$81 per cent in the case of Seyfert I versus Seyfert II.

From a methodological standpoint, the results of our experiments indicate how sensitive the performances of the photometric classification of emission line galaxies are to the size of the spectroscopic data sets used to train the method, and to the uncertainty affecting the spectroscopic classification of the training set sources.

It is important to stress that, even with a completeness of about 58 per cent, the possibility of using photometric data alone would lead to a catalogue of candidate AGN about 200 times larger than existing ones, still retaining a purity of about 70 per cent.

This work, which should be interpreted as a *feasibility study*, is hence just a first step and encourages the possibility of proceeding further with more fine classifications of the different families of line emission galaxies by exploiting their multiband photometry.

## REFERENCES

Adelman-McCarthy J. K. SDSS Consortium, 2006, ApJS, 162, 38
Alexander D. M., Vignali C., Bauer F. E., Brandt W. N., Hornschemeier A. E., Garmire G. P., Schneider D. P., 2002, AJ, 123, 1149
Annis J. T., 2013, BAAS, 221, 335.05
Antonucci R., 1993, A&A, 31, 473
Baldwin J. A., Phillips M. M., Terlevich R., 1981, PASP, 93, 5
Bishop C. M., 1995, Neural Networks for Pattern Recognition. Oxford Univ. Press, Oxford
Brescia M., Cavuoti S., d'Angelo G., D'Abrusco R., Donalek C., Deniskina N., Laurino O., Longo G., 2009, Mem. Soc. Astron. Ital. Suppl., 13, 56
Brescia M. et al., 2010, preprint (arXiv:1010.4843v2)
Brescia M., Cavuoti S., Paolillo M., Longo G., Puzia T., 2012, MNRAS, 421, 1155
Brescia M., Cavuoti S., D'Abrusco R., Longo G., Mercurio A., 2013, ApJ, 772, 140
Broyden C. G., 1970, J. Inst. Math. Appl., 6, 76
Cavuoti S., Brescia M., Longo G., Mercurio A., 2012, A&A, 546, 1
Cavuoti S., Garofalo M., Brescia M., Paolillo M., Pescape' A., Longo G., Ventre G., 2013, New Astron., 26, 12
Chang C. C., Lin C. J., 2001, Neural Comput., 13, 2119
D'Abrusco R., Staiano A., Longo G., Brescia M., Paolillo M., De Filippis E., Tagliaferri R., 2007, ApJ, 663, 752
de Jong J. T. A., Verdoes K., Gijs A., Kuijken K. H., Valentijn E. A., 2006, Exp. Astron., 35, 25
Donalek C., 2006, PhD thesis, Univ. Naples Federico II
Donley J. L., Rieke G. H., Perez-Gonzalez P. G., Rigby J. R., Alonso-Herrero A., 2007, ApJ, 660, 167
Fletcher R., 1970, Comput. J., 13, 317
Goldfarb D., 1970, Math. Comput., 24, 23
Golub G. H., Ye Q., 1999, SIAM J. Sci. Comput., 21, 1305
Hatziminaoglou E. et al., 2005, BAAS, 37, 1246
Heckman T. M., 1980, A&A, 87, 182
Kauffmann G. et al., 2003, MNRAS, 346, 1055
Kauffmann G., White S. D. M., Heckman T. M., Ménard B., Brinchmann J., Charlot S., Tremonti C., Brinkmann J., 2004, MNRAS, 353, 713
Kewley L. J., Dopita M. A., Sutherland R. S., Heisler C. A., Trevena J., 2001, ApJ, 556, 121
Kewley L. J., Groves B., Kauffmann G., Heckman T., 2006, MNRAS, 372, 961
Lamareille F., 2010, A&A, 509, A53
Mahajan S., Haines C. P., Raychaudhury S., 2010, MNRAS, 404, 1745
Marlin B. M., 2008, Missing Data Problems in Machine Learning. Library and Archives, Canada
Messias H., Afonso J., Hopkins A., Mobasher B., Dominici T., Alexander D. M., 2010, ApJ, 719, 790
Nocedal J., Wright S. J., 1999, Numerical Optimization. Springer-Verlag, Berlin
Popesso P., Biviano A., 2006, A&A, 460, L23
Seymour N., 2007, Disentangling the Evolution of Starburst and AGN Populations in Deep Radio Surveys. NOAO Proposal ID 2007A-0363
Shanno D. F., 1970, Math. Comput., 24, 647
Sorrentino G., Radovich M., Rifatto A., 2006, A&A, 451, 809
Suchkov A. A., Hanisch R. J., Margon B., 2005, ApJ, 130, 2439
Tagliaferri R., Longo G., D'Argenio B., Incoronato A., 2003a, Neural Netw., 16, 295
Tagliaferri R., Longo G., Andreon S., Capozziello S., Donalek C., Giordano G., 2003b, in Apolloni B., Marinaro M., Tagliaferri R., eds, Lecture Notes in Computer Science, Vol. 2859, Neural Networks for Photometric Redshifts Evaluation. Springer, Heidelberg, p. 226
Tonry J. L. et al., 2012, ApJ, 750, 99
Veilleux S., Osterbrock D. E., 1987, ApJS, 63, 295
Watrous R. L., 1987, in Caudill M., Butler C., eds, Proc. IEEE 1st Int. Conf. on Neural Netw., Vol. 2, Learning Algorithms for Connectionist Networks: Applied Gradient Methods of Nonlinear Optimization. IEEE Press, New York, p. 619
Winter C., Jeffery C. S., Drilling J. S., 2004, Ap&SS, 291, 375
York D. G., Adelman J., Anderson J. E. SDSS Collaboration, 2000, ApJ, 120, 1579

This paper has been typeset from a T<sub>E</sub>X/L<sup>A</sup>T<sub>E</sub>X file prepared by the author.