A&A 557, A16 (2013)
DOI: 10.1051/0004-6361/201321447

Astronomy & Astrophysics

The VIMOS Public Extragalactic Redshift Survey (VIPERS)

A support vector machine classification of galaxies, stars, and AGNs⋆

K. Małek[1,⋆⋆], A. Solarz[1], A. Pollo[2,3], A. Fritz[4], B. Garilli[4,5], M. Scodeggio[4], A. Iovino[6], B. R. Granett[6], U. Abbas[7], C. Adami[5], S. Arnouts[8,5], J. Bel[9], M. Bolzonella[10], D. Bottini[4], E. Branchini[11,12,13], A. Cappi[10], J. Coupon[14], O. Cucciati[10], I. Davidzon[10,15], G. De Lucia[16], S. de la Torre[17], P. Franzetti[4], M. Fumana[4], L. Guzzo[6,18], O. Ilbert[5], J. Krywult[19], V. Le Brun[5], O. Le Fevre[5], D. Maccagni[4], F. Marulli[15,20,10], H. J. McCracken[21], L. Paioro[4], M. Polletta[4], H. Schlagenhaufer[22,23], L. A. M. Tasca[5], R. Tojeiro[24], D. Vergani[25], A. Zanichelli[26], A. Burden[24], C. Di Porto[10], A. Marchetti[27,6], C. Marinoni[9,28], Y. Mellier[21], L. Moscardini[15,20,10], R. C. Nichol[24], J .A. Peacock[17], W. J. Percival[24], S. Phleps[23], M. Wolk[21], and G. Zamorani[10]

(Affiliations can be found after the references)

Received 11 March 2013 / Accepted 6 June 2013

ABSTRACT

Aims. The aim of this work is to develop a comprehensive method for classifying sources in large sky surveys and to apply the techniques to the VIMOS Public Extragalactic Redshift Survey (VIPERS). Using the optical (u^*, g', r', i') and near-infrared (NIR) data (z', K_s), we develop a classifier, based on broad-band photometry, for identifying stars, active galactic nuclei (AGNs), and galaxies, thereby improving the purity of the VIPERS sample.

Methods. Support vector machine (SVM) supervised learning algorithms allow the automatic classification of objects into two or more classes based on a multidimensional parameter space. In this work, we tailored the SVM to classifying stars, AGNs, and galaxies and applied this classification to the VIPERS data. We trained the SVM using spectroscopically confirmed sources from the VIPERS and VVDS surveys.

Results. We tested two SVM classifiers and concluded that including NIR data can significantly improve the efficiency of the classifier. The self-check of the best optical + NIR classifier has shown 97% accuracy in the classification of galaxies, 97% for stars, and 95% for AGNs in the 5-dimensional colour space. In the test of VIPERS sources with 99% redshift confidence, the classifier gives an accuracy equal to 94% for galaxies, 93% for stars, and 82% for AGNs. The method was applied to sources with low-quality spectra to verify their classification, hence increasing the security of measurements for almost 4900 objects.

Conclusions. We conclude that the SVM algorithm trained on a carefully selected sample of galaxies, AGNs, and stars outperforms simple colour–colour selection methods and can be regarded as a very efficient classification method particularly suitable for modern large surveys.

Key words. methods: data analysis – methods: statistical – surveys – galaxies: fundamental parameters – stars: fundamental parameters – cosmology: observations

1. Introduction

Over the years, the amount of astronomical data collected by satellites and ground-based surveys is steadily increasing. The zoo of collected data, such as photometry, redshifts, spectral lines, and morphology, is constantly expanding, and increasingly researchers are turning to automated algorithms to explore the high-dimensional parameter space. Although computationally challenging, the goal is to make use of every available feature to recognise and extract the most discriminating patterns and allow full systematisation of the data.

Furthermore, the study of the dependence of galaxy properties on physical parameters such as galaxy mass or environment can greatly benefit from the efficient classification of sources. The classification of different types of sources is one of the basic and, at the same time, crucial tasks to perform before moving on to any scientific analysis.

The first physical classification of sources in a photometric sky survey is between foreground stars within the Galaxy and extragalactic sources. Generally, the distinction between stars and galaxies can be made based upon morphological measurements; point sources are classified as stars, while extended sources are classified as galaxies (e.g. Vasconcellos et al. 2011; Henrion et al. 2011). For bright apparent magnitudes, the morphology appears to be a reliable criterion for classifying stars and galaxies, but at fainter magnitudes it becomes difficult to detect

⋆ Based on observations collected at the European Southern Observatory, Cerro Paranal, Chile, using the Very Large Telescope under programme 182.A-0886 and partly 070.A-9007. Also based on observations obtained with MegaPrime/MegaCam, a joint project of CFHT and CEA/DAPNIA, at the Canada-France-Hawaii Telescope (CFHT), which is operated by the National Research Council (NRC) of Canada, the Institut National des Sciences de l'Univers of the Centre National de la Recherche Scientifique (CNRS) of France, and the University of Hawaii. This work is based in part on data products produced at TERAPIX and the Canadian Astronomy Data Centre as part of the Canada-France-Hawaii Telescope Legacy Survey, a collaborative project of NRC and CNRS. The VIPERS web site is http://www.vipers.inaf.it/

⋆⋆ Postdoctoral Fellow of the Japan Society for the Promotion of Science.

low-brightness objects like ultra-compact dwarf (UCD) galaxies, which are often misclassified as foreground stars (Drinkwater et al. 2003). Resolved stellar selection in the current and next generation of wide-field surveys, such as Euclid (Laureijs et al. 2012), BigBOSS (Sholl et al. 2012), DES (Mohr et al. 2012), LSST (Ivezic et al. 2009), LAMOST (Bland-Hawthorn 2012), and Pan-STARRS (Kaiser et al. 2010), and/or deep surveys, such as VUDS[1] (Lefevre et al., in prep.), HUDF (Beckwith et al. 2006), DLS (Wittman et al. 2002), and VISTA (Emerson & Sutherland 2010), is being challenged by the vast number of unresolved galaxies at faint apparent magnitudes (Fadely et al. 2012). Including near-infrared (NIR) photometric bands for many new surveys should improve the classification and separation of faint sources and stars, thereby providing an alternative method of spectroscopy.

In the case of fainter sources, colour–colour diagrams are the most widely used tools to separate different classes of celestial sources from one another, since different types of objects will appear in different colour regions in such diagrams due to the shape of the spectral energy distribution (SED). For example, galaxies possess much redder colours than do stars owing to the higher flux at longer wavelengths (e.g., Walker et al. 1989). Classification methods based on colour–colour selection were employed for star-galaxy separation (e.g. infrared colour diagram used by Pollo et al. 2010) or for finding special classes of sources, such as high/low-redshift quasars, active galactic nuclei (AGNs), starburst galaxies, or variable stars (Richards et al. 2002; Stern et al. 2005, 2012; Chiu et al. 2005; Brightman & Nandra 2012; Woźniak et al. 2004).

Support vector machines (SVMs) are a class of supervised learning algorithms that were created as an extension to nonlinear models of the generalised portrait algorithm developed by Vladimir Vapnik (Vapnik 1995), for classification in a multidimensional parameter space. These algorithms are based on the concept of decision planes to classify objects using their relative positions in the n-dimensional parameter space. A large number of observed properties may be analysed simultaneously by the classifier making full use of the data. Within the full parameter space, it is possible to build a more reliable classifier than is possible by only using a subset of the data (for example, by analysing only two photometric colours, instead of the complete set). On the other hand, the method requires a training sample, that is, a set of data that have known classifications. Generally, SVM algorithms are sensitive to the measurement errors and are of limited use for extracting information from noisy data sets (Fadely et al. 2012). The classification of observed sources in astronomy is a fundamental problem, and there is still no approach completely free of drawbacks; however, SVM algorithms are a novel and very promising classification strategy.

In this paper we apply the SVM algorithm to photometric data. Previous works (e.g., Fadely et al. 2012; Solarz et al. 2012; Vasconcellos et al. 2011; Ball et al. 2006) show high efficiency in that approach for two classes of objects (galaxies and stars). Recently, the Photometric Classification Server (PCS) for the prototype of the Panoramic Survey Telescope and Rapid Response System (Pan-STARRS1) based on SVMs was developed (Saglia et al. 2012). The PCS system is using five photometric bands (g_{P1}, r_{P1}, gi_{P1}, z_{P1}, and y_{P1}) and is able to separate three groups of sources (stars, galaxies, QSOs) without any preselection based on colour or redshift range and with high accuracy of galaxy classification (~97%). The purities of stellar and QSO samples' classifications are worse, at the levels of 85% and 83%, respectively.

We decided to develop a three-class recognition algorithm, which will be able to classify galaxies/AGNs/stars based on the photometric data in The Canada France Hawaii Telescope Legacy Survey (CFHTLS). We used, as a training set in colour space objects with the best-quality spectra from the VIMOS Public Extragalactic Redshift Survey (VIPERS) and VIMOS VLT Deep Survey (VVDS) Deep (F02 field) and Wide (F22 field) data. After carefully selecting objects from VIPERS by SVM and defining characteristic patterns for different types of sources, it will be possible to enlarge the sample of galaxies to be used for more detailed studies. We plan to use this trained classifier on a large number of sources possessing low-quality spectra within VIPERS to recover sources that cannot be classified based upon the spectrum alone. A majority of objects with lower quality spectral information are absorption line systems with low signal-to-noise ratio. Faint red stars and faint passive galaxies are often difficult to distinguish by their spectral features, if the quality of a spectrum is low. Reconfirmation of a class of such an object by the SVM classifier (galaxy, AGN, or star) based upon the photometric measurements also increases the probability that their spectroscopically measured redshift is correct.

The paper is organised as follows. In Sect. 2, we describe the data used in our analysis, both spectroscopic and photometric. Section 3 describes the principles of the SVM learning algorithm. In Sect. 4 we introduce the training sample used in our work. In Sect. 5, we compare the efficiency of the classifier with and without near infrared data. Additionally, we present the results of the analysis of the basic tests for the classifiers – self-check and test of the classifier on the VIPERS galaxies with redshift measurements confirmation level equal to 95%. The section closes with the selection of the optimal classifier used for our subsequent analysis. Section 6 describes the results of our classification of optical NIR SVM classifier objects from the VIPERS samples. Finally, in Sect. 8 we discuss the advantages and limitations of our current SVM classifier, and we outline our improvements for the presented classifier.

2. Data

2.1. Photometric data

In this section we present the photometric data used in our work. All quoted magnitudes used to develop SVM classifiers are in the AB photometry system and were corrected for foreground Galactic extinction according to the $E(B-V)$ factor derived from Schlegel maps (Schlegel et. al. 1998). The correction for Galactic extinction was performed for each source separately (see Fritz et al. 2013). The mean value of $E(B-V)$ factor for the CFHTLS W1 field is equal to 0.02 mag, and for the CFHTL W4 field it is equal to 0.05 mag.

CFHTLS photometry

The CFHTLS, a joint Canadian-French programme, has three distinct survey components: (1) the SuperNovae Legacy Survey the "Deep" survey; (2) the "Wide" – wide synoptic survey (on which VIPERS survey was based); and (3) a very wide shallow survey, the "Very Wide".

The heart of MegaPrime, the wide-field optical imaging facility, is the MegaCam CCD camera (Boulade et al. 2000). MegaCam provides multicolour photometry with wavelength (λ)

[1] http://cesam.oamp.fr/vuds

Table 1. MegaPrime* and WIRCam** filter characteristics.

Filter	u^*	g'	r'	i'	z'	K_s
Central λ (nm)	374	487	628	777	1170	2146
Bandwidth (nm)	76	145	122	151	687	325
Max. transmission (%)	77.5	93.5	96.3	98	95	98
Mag. limit***	25.30	25.50	24.80	24.48	23.60	22.00

Notes. (*) http://www.cfht.hawaii.edu/Instruments/Filters/megaprime.html (**) http://www.cfht.hawaii.edu/Instruments/Filters/wircam.html (***) Measured as the 50% of completeness (MegaPrime) and 5σ (WIRCam) for point sources.

coverage from 3500 to 9400 Å. The main characteristics of the MegaPrime/MegaCam broad band filters are described in Table 1. For a more detailed description we refer the reader to the CFHTLS[2] official web page.

The data used in this work are a part of CFHTLS T0005 release (Mellier et al. 2008), produced at the TERAPIX[3] data centre. We consider a subsample of CFHTLS T0005 catalogue with spectroscopic redshift measured by VIPERS.

The CFHTLS data are provided in single tiles with effective area of ∼1 deg square, which partially overlap each other. During the preparation of the input data for spectroscopic observations we found the shift in colours between different tiles. To obtain a homogeneous colour selection of spectroscopic targets, the tile-to-tile correction was performed by using one of the fields overlapping with the VVDS-Deep survey (W1-25) as a representative tile. The detailed description of the tile-to-tile correction and the explanation of the colour correction method can be found in the survey description paper (Guzzo et al. 2013).

WIRCam data

In our work, we also used NIR K_s measurements in the AB magnitude system, which were corrected for galaxy extinction and taken from Wide-field InfraRed Camera (WIRCam; Thibault et al. 2003; Puget et al. 2004), coming from the dedicated follow-up observations for the VIPERS project (Arnouts et al., in prep.). The K_s filter has a central wavelength of 2146 nm, and maximum transmission on the level of 98%. One may find the detailed description of WIRCam detector on the WIRCam CFHT web page[4].

2.2. Spectroscopic data

VIPERS survey

The VIMOS Public Extragalactic Redshift Survey[5] is an ongoing large programme aimed at measuring redshifts for $\sim 10^5$ galaxies at redshift $0.5 < z \lesssim 1.2$, to accurately and robustly measure clustering, the growth of structure (through redshift-space distortions), and galaxy properties at an epoch when the Universe was about half its current age. The galaxy target sample is selected from optical photometric catalogues of the Canada-France-Hawaii Telescope Legacy Survey Wide (CFHTLS-Wide, Goranova et al. 2009; Mellier et al. 2008). VIPERS covers ∼24 deg^2 on the sky and is divided into two areas within the W1 and W4 CFHTLS fields. Galaxies are selected to a limit of $i_{AB} < 22.5$ measured using Sextractor's mag_auto (Kron 1980)-like magnitude. In addition, a simple and robust colour preselection in $(g - r)$ vs. $(r - i)$ is applied to efficiently remove galaxies at $z < 0.5$. In combination with an efficient observing strategy (Scodeggio et al. 2009), this allows us to double the galaxy sampling rate in the redshift range of interest with respect to a purely magnitude-limited sample, reaching an average target sampling rate of >40%. At the same time, the area and depth of the survey results in a fairly large volume, 5×10^7 h^{-3} Mpc^3, analogous to that of the 2dFGRS at $z \sim 0.1$ (Colless et al. 2001, 2003). This combination of sampling and depth is quite unique over current redshift surveys at $z > 0.5$.

VIPERS spectra are collected with the VIsible imaging Multi-Object Spectrograph (VIMOS, Le Fèvre et al. 2000) at moderate resolution ($R = 210$), using the LR red grism, providing a wavelength coverage of 5500–9500 Å, for a typical redshift rms error of $\sigma_z = 0.00047\ (1 + z)$. The full VIPERS area of ∼24 deg^2 is covered through a mosaic of 288 VIMOS pointings (192 in the W1 area, and 96 in the W4 area). Of the VIPERS spectroscopic targets, more than 51 000 K_s counterparts were found: 96% (80%) of our spectra for W1 (W4) field have K_s measurements. More detailed description of WIRCam follow-up survey for VIPERS project can be found in Fritz et al. (2013) and Davidzon et al. (2013).

The redshift quality is quantified at the time of validation by attributing grading flags ($\mathrm{VIPERS}_{\mathrm{Zflag}}$) that are obtained from repeated measurements of redshift for the same sources. The $\mathrm{VIPERS}_{\mathrm{Zflag}}$ for galaxies and stars range from a value of 4, indicating >99% of confidence that the measurement is secure, to 0, representing a lack of a reliable estimate of redshift. $\mathrm{VIPERS}_{\mathrm{Zflag}}$ equal to nine corresponding to galaxies with only one single clear spectral emission feature. Objects classified as AGNs follow the same scheme but their flags are increased by ten. A similar system was used and tested for example for VVDS survey (Le Fèvre et al. 2005). A discussion of the survey data reduction and management infrastructure is presented in Garilli et al. (2012). An early subset of the spectra used here has been analysed and classified through a principal component analysis (PCA) in Marchetti et al. (2012). A more complete description of the survey construction, from the definition of the target sample to the actual spectra and redshift measurements, is given in the parallel survey description paper, Guzzo et al. (2013).

The data set used in this paper are those of the early science data release of VIPERS data as described in Guzzo et al. (2013); see also de la Torre et al. (2013), Fritz et al. (2013), Marulli et al. (2013), Bel et al. (2013), and Davidzon et al. (2013). This data will be publicly available in fall 2013 as the VIPERS Public Data Release 1 (PDR-1) catalogue. This catalogue includes 55 358 redshifts and corresponds to the reduced data as it was in the VIPERS database at the end of the 2011/2012 observing campaign.

Using the automatic source classifier for VIPERS data is a natural step to handle this unique data volume. Automated and efficient source classifiers based on photometric observations, can provide class labels for catalogues and be used to recover objects for study according to various criteria. Moreover, a multilevel SVM classifier, trained to search for specific types of sources such as AGNs or galaxies, with an additional redshift measurement as a feature in the parameter space, can be used to boost confidence in the reliability of redshift estimates for sources with poor spectroscopic data. We are planning to develop a more sophisticated and detailed classifier in the near

[2] http://www.cfht.hawaii.edu/Science/CFHTLS/
[3] http://terapix.iap.fr/
[4] http://www.cfht.hawaii.edu/Instruments/Imaging/WIRCam/
[5] See http://vipers.inaf.it

future, enlarging the parameter space by adding measurements of spectral lines and galaxy morphological parameters, thus enabling a finer classification of our sources (e.g. distinguish among different galaxy types).

In this work, we used VIPERS data both to construct a training sample and to select samples on which to apply the classifier to separate three different classes of objects (galaxy/AGN/star).

VIMOS-VLT Deep Survey (VVDS)

VIPERS was designed as an extragalactic survey that aims to efficiently measure of redshifts for a large sample of galaxies. To increase the efficiency, stars were carefully removed from the target candidates (which was particularly important for the W4 VIPERS field owing to its low galactic latitude). To this aim, both morphological and SED fitting techniques were used (see Guzzo et al. 2013; Coupon et al. 2009). However, it was also important to re-introduce AGNs, which were identified among the stellar objects by their photometric properties (a more detailed description of AGN selection can be found in the survey description paper, Guzzo et al. 2013). Consequently, the number of observed stars and AGNs in VIPERS is quite small.

To construct a reliable training sample (see Sect. 3), we included data from another, similar, but more complete survey, VVDS. The VVDS fields, like VIPERS, are covered by CFHTLS (and partially by WIRCam observations) and thus the photometric information is homogeneous. Additionally, both surveys utilise the VIMOS spectrograph in similar configurations. The VVDS spectroscopic sample is based upon a purely magnitude-limited selection such that the survey contains a much wider variety of sources than VIPERS. We used VVDS-Deep (F02 field) and VVDS-Wide (F22 field) surveys to construct a training sample of AGNs (objects classified as AGNs by Gavignaud et al. 2007). The stellar sample was chosen from a part of VVDS Wide F22 that overlaps the VIPERS W4 field.

The Deep F02 survey, covering 0.49 square degrees, is a purely magnitude limited sample to $i_{\rm AB} \leq 24$. The detailed description of the VVDS Deep survey may be found in Le Fèvre et al. (2005). The VVDS Wide F22 survey (Garilli et al. 2008), covering an effective area three square degrees, is also a magnitude limited survey with limitation to $i_{\rm AB} = 22.5$.

3. Method – support vector machines

The main purpose of the SVM is to calculate decision planes between a set of objects having different class memberships. A so-called training sample, a training set of objects, is used to provide the SVM with examples of the different classes of sources. The SVM searches for the optimal separating hyperplane between the n different classes of objects by maximising the margin between the classes closest points (the so-called support vectors). Instead of using the probability function as in Bayesian statistics or template-fitting methods, the objects are classified based on their relative position in the n-dimensional parameter space with respect to the separation boundary. A well chosen training sample is at the heart of the method, because, based on the properties of the training sample, the classifier is tuned, and the hyperspace between classes is determined.

The SVM algorithm represents a major development in machine-learning techniques. It can be applied to classification or regression problems and is nowadays constantly growing in popularity, to deal with astronomical data for distinguishing different classes of sources based on a multidimensional space of parameters taken from observations. Recently, Woźniak et al. (2004) has used SVMs efficiently to analyse variable sources in a five-dimensional space constructed from the period, amplitude, and three colours. Huertas-Company et al. (2008) quantified the morphologies of NIR galaxies based on 12-dimensional space, including five morphological parameters and other characteristics of galaxies, such as luminosity and redshift. Solarz et al. (2012) created a star-galaxy separation algorithm based on mid and NIR colours, and Saglia et al. (2012) separated three different classes of sources (galaxies, QSOs, and stars) from the PAN-STARRS1 survey, based on five photometric bands. Last year brought a significant number of astronomical papers that implement supervised machine-learning algorithms to handle various tasks, not only to classify sources but also to predict characteristic features of specific objects. For example, Peng et al. (2012) used SVM to select AGN candidates and to estimate redshift, Hassan et al. (2013) – to search specific AGN subclass: BL Lacertae and flat-spectrum radio quasars based on the Second *Fermi* LAT Catalogue). Clearly SVMs present an innovative method with great potential to be widely used in many different branches of astronomy, a potential we are just beginning to tap into.

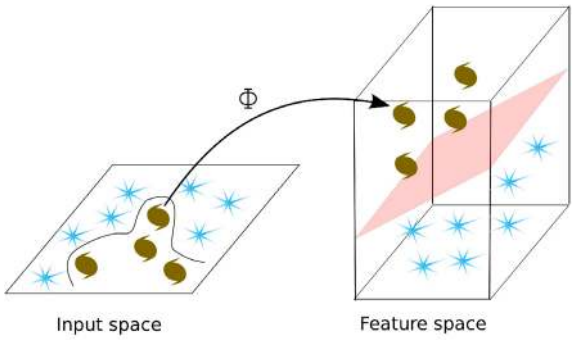

Fig. 1. Illustration of the operation of the SVM algorithm. The input data (on the *left side*) are transformed by a kernel into the higher dimensional feature space (*right side*) where, instead of having a complex boundary separating different classes of objects, we can find an optimal separating hyperplane.

We used the SVM algorithm to build a non-linear classifier for photometric data to select three different classes of objects: galaxies, AGNs, and stars. The first step in our classification task involves selecting a secure training sample of galaxies, AGNs, and stars, taking advantage of the redshift information provided by VIPERS and VVDS and using their attributes – i.e. their observed photometric fluxes – to train the SVM.

The algorithm, aided by a non-linear kernel function, searches for a hyperplane that will maximise the distance from the boundary to the closest points belonging to the separate classes of objects (Cristianini & Shawe-Taylor 2000; Shawe-Taylor & Cristianini 2004). The kernel is a symmetric function Φ that maps $k : X \times X \rightarrow F$, so that for all x_i and x_j, $k(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle$ from the input space X to the feature space F (Vanschoenwinkel & Manderick 2005), see Fig. 1. For our analysis we chose a Gaussian radial basis kernel (RBK) function, defined as

$$k(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right), \quad (1)$$

where $\|x_i - x_j\|$ is the Euclidean distance between x_i, and x_j. The effect of the kernel function is a non-linear representation of each parameter from the input to the feature space. The RBK kernel is one of the most popular SVM kernel functions, used to make

the non-linear feature map. We decided to use it because of its effectiveness and simplicity in adjusting the free parameters.

For our tasks, we used a soft-boundary SVM method called C-SVM. We chose C-classification because of its good performance and only two free parameters:

- C – a trade-off parameter that sets the width of the margin separating different classes of objects. A large C value sets a small margin of separation between different classes of objects; however increasing the C parameter too much can lead to over-fitting. Reducing C will make the hyperplane between different classes of objects smoother, allowing for some misclassifications.
- $\gamma > 0$ parameter (related to the kernel function) determines the topology of the decision surface. A low value of γ sets a very rigid, and complicated decision boundary; a value of γ that is too high can give a very smooth decision surface causing misclassifications.

A schematic representation of the SVM algorithm classification process, beginning with choosing the training sample, tuning C and γ parameters, self-checking of the classifier, and finally, classifying the real sample is shown in Fig. 2.

For our analysis we used LIBSVM[6] (Chang & Lin 2011), an integrated software for support vector classification, which allows for multiclass classification. We used R[7], a free software environment for statistical computing and graphics, with e1071 interface (Meyer 2001) package installed.

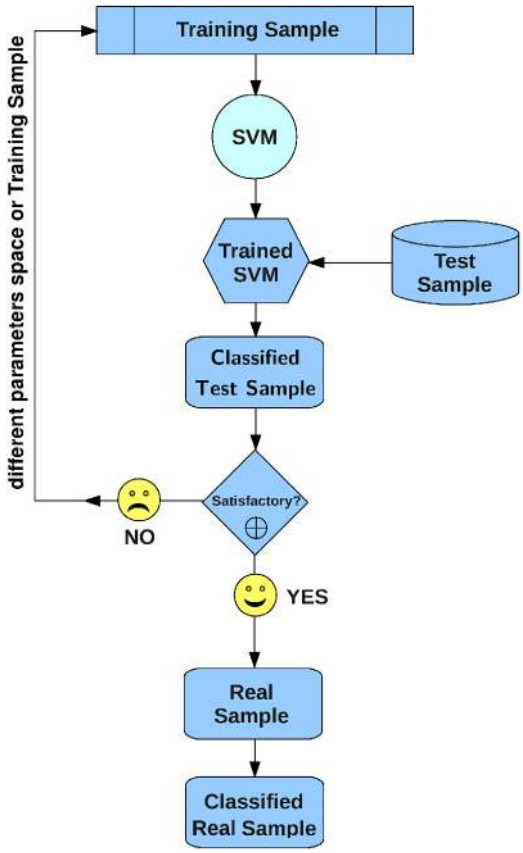

Fig. 2. Schematic representation of the SVM algorithm classification process. We take as input the preselected training sample consisting of (in the case of this work) three distinct classes of objects. The SVM is taught how to distinguish one class from the others based on the discriminating properties chosen as feature vectors. Then, the classifier is trained by tuning the free parameters (C and γ). If the result reaches a high enough accuracy rate (the number of objects from the training sample that are correctly recognised by the classifier) without overfitting (the resulting hyperplane does not confine the sources of a specific type too tightly), it will be used to classify the unknown objects (test sample). If the accuracy is not satisfactory, a different parameter space (or training sample, if possible) is chosen to tune C and γ. After a number of iterations, which allow the classifier to reach high enough efficiency level, a real sample can be classified using the discriminant hyperplanes.

4. Training sample

The successful application of an SVM algorithm requires a carefully selected training sample – a set of objects with confirmed classes which will serve as a template for distinguishing the sources whose class we want to determine. Since this work is focused on the selection of galaxies, AGNs, and stars we select as a training sample a set of sources whose basic class (galaxy, AGN or star) was established with the highest reliability thanks to their high quality spectra (their redshift being measured with the highest confidence flag within the VIPERS or VVDS surveys). For these sources the accurate photometric information provided by the CFHTLS wide-survey and the WIRCam follow-up observations of the VIPERS/VVDS fields, provided the colour information needed to create the discriminant vectors for training our SVM algorithm. We produced a model (the optimised C and γ parameters based on the training data), which predicts the target values of the test data given only the test data attributes (Hsu et al. 2010).

4.1. Galaxies

As a galaxy training sample we used the sources with the best redshift measurements in both the W1 and W4 VIPERS fields ($\mathrm{VIPERS}_{\mathrm{Zflag}} = 4$, corresponding to the highest confidence level of redshift measurements and thus of spectroscopic classification as a galaxy). It is useful to remember that VIPERS is preselected not only in magnitude ($i' < 22.5$) but also in colours: $(r' - i') > 0.5 * (u^* - g)$ or $(r' - i') > 0.7$. We have divided the galaxy training set into i'-based apparent magnitude-binned samples and trained the classifier on each subset. As a galaxy training sample we used 16 271 galaxies: 1884, 5483, 6778, and 3226 for $19 \leqslant i' < 20$, $20 \leqslant i' < 21$, $21 \leqslant i' < 22$, and $22 \leqslant i' < 22.5$ apparent magnitude-bins, respectively. Based on our initial tests, we decided to divide our galaxy sample into the magnitude bins to separate more efficiently different groups of galaxies seen in different i' apparent magnitude ranges to improve their classification. Figure 3 shows that galaxies in different magnitude bins occupy different areas of the colour–colour plots, partly because of different redshift range and different morphology.

4.2. AGNs

Given the small number of AGNs detected in the VIPERS fields with the $\mathrm{VIPERS}_{\mathrm{Zflag}} = 14$, we increased the AGN sample by using all AGNs which had at least 99% confidence level of spectroscopic classification ($\mathrm{VIPERS}_{\mathrm{Zflag}}$ 13 and 14, in total 398 objects). AGN spectra are quite easy to recognise, so a lower flag on the quality of the measured redshift does not infringe on the

[6] http://www.csie.ntu.edu.tw/cjlin/~libsvm/
[7] http://www.r-project.org/

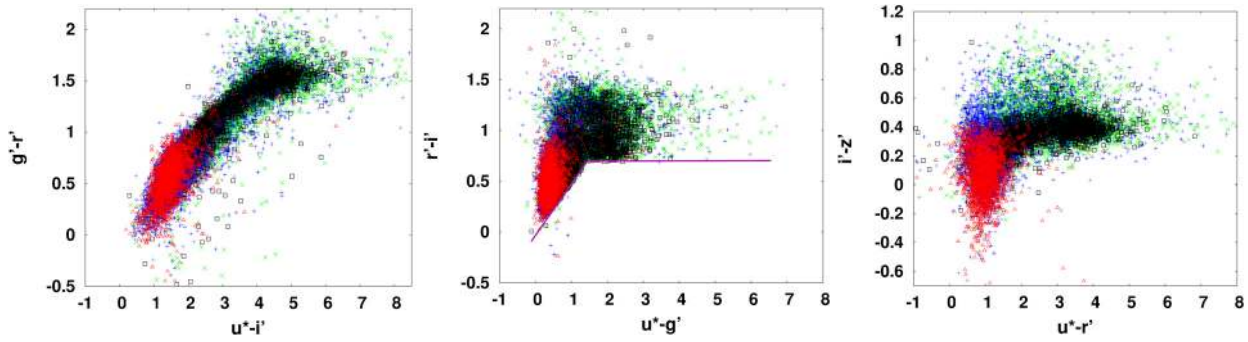

Fig. 3. Representative colour–colour plots for the galaxy training sample. Open black squares represent objects with i'-apparent magnitude between 19 and 20 mag; green X-s – galaxies with i' magnitude between 20 and 21 mag; objects with i' apparent magnitude between $21 \leqslant i < 22$, and $22 \leqslant i < 22.5$ mag are marked as blue +-s and open red triangles, respectively; in the *middle panel* of colour–colour plots, the boundaries of VIPERS selection are marked as magenta lines.

reliability of the classification as an AGN. There are two ways that an AGN can be observed in VIPERS:

- it is star-like and meets the AGN candidate selection. This includes samples of X-ray selected AGNs from the XMM-LSS survey, overlapping the VIPERS W1 field (Pierre et al. 2004), and AGNs selected by colour–colour criteria from the sample of star-like sources that would otherwise not be targeted.
- it meets the galaxy selection criteria – AGNs which met the galaxy criteria during the main VIPERS colour preselection.

We stress that the colour preselection for galaxies and AGNs is slightly different, and AGNs occupy only a part of the full colour–colour galaxy plane. The first AGN colour separation criterion $\mathrm{CC}_{1\mathrm{AGN}}$:

$$(g'-r') < 1 \wedge \begin{cases} 1.\ (u^*-g)_{\mathrm{corr}} < 0.6, \\ 2.\ \begin{array}{l} 0.6 \leqslant (u^*-g)_{\mathrm{corr}} < 1.2 \text{ and} \\ (g'-r')_{\mathrm{corr}} > 0.5(u^*-g')_{\mathrm{corr}} + 0.036, \end{array} \\ 3.\ \begin{array}{l} 0.6 \leqslant (u^*-g)_{\mathrm{corr}} < 2.6 \text{ and} \\ (g'-r')_{\mathrm{corr}} < 0.5(u^*-g')_{\mathrm{corr}} + 0.214, \end{array} \\ 4.\ (u^*-g')_{\mathrm{corr}} > 2.6, \end{cases} \quad (2)$$

where $(u^*-g)_{\mathrm{corr}}$ and $(g'-r')_{\mathrm{corr}}$ correspond to tile colour offset.

The colour–colour selection criterion of AGNs, given in Eq. (2), was based on the results from the VVDS survey. After one year of observations it turned out that this selection criterion introduces a stellar contamination at the level ~60%. From August 2010, additional criterion $\mathrm{CC}_{2\mathrm{AGN}}$, including the $(g'-i')$ vs. (u^*-g') colour–colour plane, was added to eliminate stellar sample from AGNs targets. The set of colour–colour criteria included to $\mathrm{CC}_{2\mathrm{AGN}}$ is

$$\begin{cases} 1.\ (u^*-g')_{\mathrm{corr}} < 0.6 \text{ and } -0.2 < (g'-i') < 1, \\ 2.\ \begin{array}{l} 0.6 \leqslant (u^*-g')_{\mathrm{corr}} < 1 \text{ and} \\ -0.2 < (g'-i') < 0.2, \end{array} \\ 3.\ (u^*-g')_{\mathrm{corr}} \geqslant 1 \text{ and } (g'-i') < 0.6. \end{cases} \quad (3)$$

Therefore, both criteria (Eqs. (2) and (3)) applied simultaneously defined VIPERS AGN targets. However, most of the AGNs share the same colour–colour space as galaxies (as can be seen in Fig. 6). A part of AGNs occupy different colour–colour areas than galaxies and for them, the galaxy/AGN separation is not so difficult. For objects classified as AGNs lying in the same colour–colour plane, the galaxy/AGN/star separation is more challenging. For this reason we decided to use SVM with n-dimensional photometric parameter space to classify sources with similar properties in the typical colour–colour plane. That is why it is a challenge to distinguish all three classes of objects using an automatic classifier.

To enlarge the AGN training sample, we also merged the VIPERS sample with objects classified as broad-line AGNs in the VVDS survey. In our training sample we included AGNs identified by Gavignaud et al. (2007) – a catalogue of broad emission-line AGNs, from the purely flux-limited spectroscopic sample of the VVDS survey. No colour-based preselection has been applied to these AGNs. For our studies we used 100 AGNs from VVDS Deep F02 (Le Fèvre et al. 2005) and VVDS Wide F22 (Garilli et al. 2008) fields only. We selected these fields since they have the same CFHTLS photometry system as the VIPERS survey. We found that AGNs detected in both VIPERS fields do not display any systematic difference in the colour–colour distribution, confirming that our extinction correction works well.

Cumulatively, our AGN training sample reached 498 objects. A part of them, observed by VIPERS, preselected by colour. AGNs from VVDS fields have no colour preselection (flux-limited only). Since we checked on colour–colour plots (see Fig. 4), in the different magnitude bins, we do not see a change in population of our AGN sample with apparent luminosity. For this reason, unlike the case of the galaxy sample, we decided not to divide the AGN training sample into i'-based apparent magnitude binned samples, but to use it as a whole in each bin to increase the population of the training AGNs.

4.3. Stars

VIPERS performed a star/galaxy classification in the CFHTLS wide fields to effectively remove stars from the sample of observed targets. This procedure is crucial, since at $i' < 22.5$ the fraction of stars can be as high as 50% (as in the case of W4, Guzzo et al. 2013). The basic VIPERS classification procedure was based on the colour–colour preselection with $(r-i) > 0.5 * (u-g)$ or $(r-i) > 0.7$, but owing to the low galactic latitude of W4 field, VIPERS implemented an additional procedure. We refer the interested reader to Guzzo et al. (2013) for a complete description of the adopted strategy, but here it is sufficient to mention that for objects brighter than $i' = 21$ an additional preselection based on the observed angular size of sources was applied, while for objects fainter than $i' = 21$ a combined method making use of an angular size and SED fitting by the Le Phare code (Arnouts et al. 1999; Ilbert et al. 2006) has been used. These preselection criteria proved to be

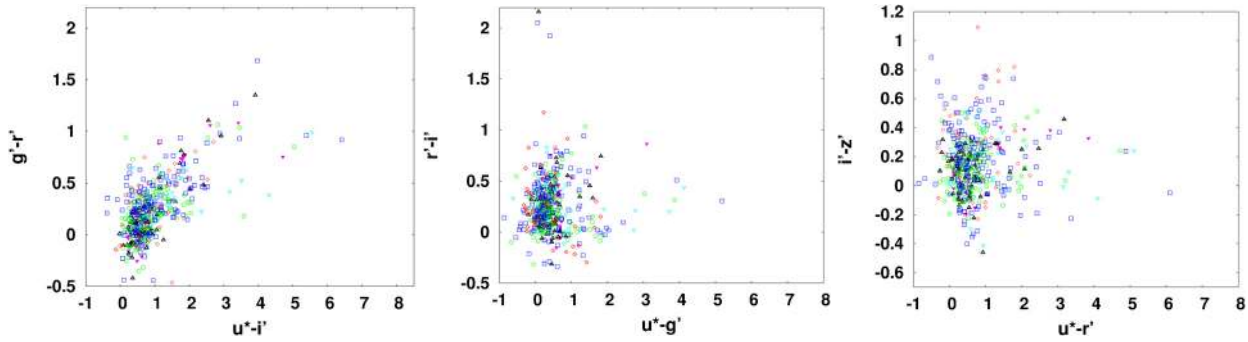

Fig. 4. Representative colour–colour plots for the AGN training sample. Full magenta triangles represent objects brighter than 19 mag in the i' band. Open black triangles – AGNs with i-apparent magnitude between 19 and 20 mag; open green circles – AGNs with i' magnitude between 20 and 21 mag; objects with i' apparent magnitude between $21 \leqslant i < 22$, and $22 \leqslant i < 22.5$ mag are marked as open blue squares and open red diamonds, respectively; AGNs with i' apparent magnitude fainter than 22.5 are marked as open rotated cyan triangles.

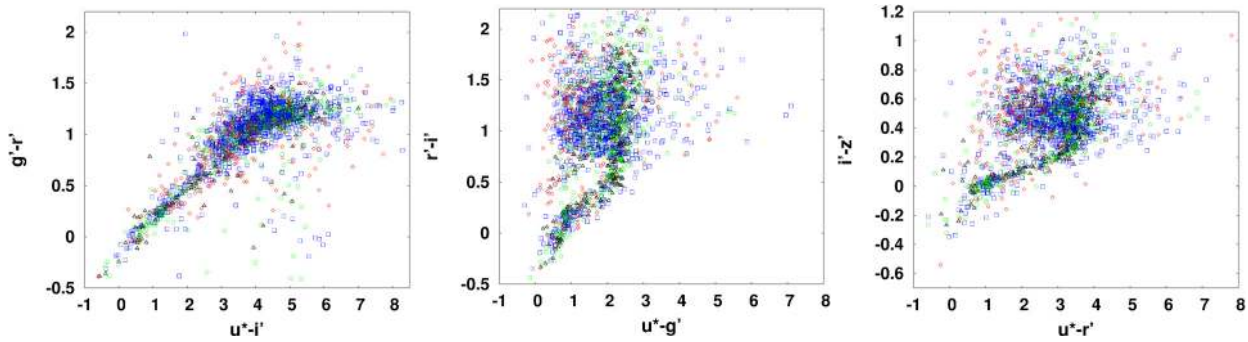

Fig. 5. Representative colour–colour plots for the star training sample. Open black triangles – stars with i-apparent magnitude between 19 and 20 mag; open green circles – stars with i' magnitude between 20 and 21 mag; objects with i' apparent magnitude between $21 \leqslant i < 22$, and $22 \leqslant i < 22.5$ mag are marked as open blue squares and open red diamonds, respectively.

very effective. However, the average stellar contamination in the VIPERS database, for both fields, remains on the level of 3.2% (1.49% and 4.86% for the W1 and W4 fields, respectively). It means that in the VIPERS PDR-1 catalogue, which includes 55 358 objects, only 1750 objects have been identified as stars. In sum, the VIPERS PDR-1 catalogue contains 1750 (3.20%) stars classified as galaxies in the beginning, with colours compatible with an object at $z > 0.5$. This stellar sample can be divided into two main groups:

- stars that were not distinguishable from galaxies based on the VIPERS preselection criteria; and
- stars that were included in the sample as AGN candidates.

Then, it should be stressed that the stars observed by VIPERS are interlopers within the galaxy and AGN samples and are thus not representative of the stellar class. However, our method uses the multidimensional colour space which opens a possibility that in such a space, these sources may occupy a region separated from galaxies and AGNs.

To build an unbiased star training sample we added spectroscopically classified stars from the VVDS Wide F22 overlap with the VIPERS W4 field. VVDS Wide F22 observations were carried out on the same magnitude limits sample as VIPERS, but without any photometric preselection. The overlap between the VVDS Wide F22 and VIPERS W4 fields contains 920 objects spectroscopically classified as stars by VVDS in the $19 \leqslant i' < 22.5$ apparent magnitude bin. We increased the stellar training sample by using all VIPERS stars with $\mathrm{VIPERS}_{\mathrm{Zflag}}$ equal to 4, in the same apparent magnitude bin (1312 objects). Cumulatively, our stellar training sample reached 2232 objects.

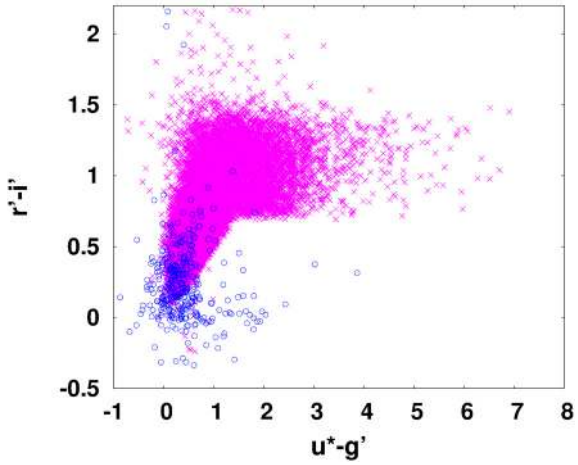

Fig. 6. Representative colour–colour plot for VIPERS galaxies with $\mathrm{VIPERS}_{\mathrm{Zflag}} = 4$ (pink x-s) and AGNs with $\mathrm{VIPERS}_{\mathrm{Zflag}} = 3$ and 4 (open blue circles).

Similar to the case of the AGN training sample, we did not divide the stellar training sample in i'-based apparent magnitude bins. As shown on the representative colour–colour plots for the different magnitude i' bins (Fig. 5), we did not observe a significant change in the distribution of our stellar sample as a function of apparent luminosity.

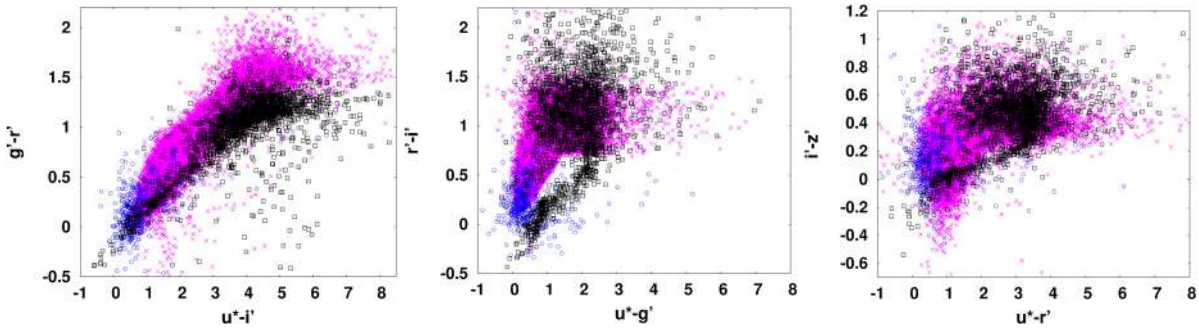

Fig. 7. Representative colour–colour plots for all objects used for the training sample. Pink x-s represent galaxies. Open blue circles correspond to the AGN sample, and open black squares to the stellar sample.

Table 2. Number (N) of galaxies, AGNs, and stars in our training sample after using the oversampling method.

	$19 \leqslant i' < 20$	$20 \leqslant i' < 21$	$21 \leqslant i' < 22$	$22 \leqslant i' < 22.5$
N galaxies	1884	5483	6778	2126
N AGNs	1520	4440	5440	1760
N stars	2232	4440	5 440	2232

4.4. Oversampling

Our training sample includes more than 16 000 galaxies, and only 2232 stars and 498 AGNs. Figure 7 shows the representative colour–colour plots for galaxies, AGNs, and stars chosen for the best training sample set. Sampling strategies, such as oversampling and undersampling, are popular solutions for tackling the problem of classification because the SVM classifier is sensitive to a high-class imbalance, resulting in a drop in the classification performance (e.g., Tang et al. 2009; Akbani et al. 2004; Raskutti & Kowalczyk 2004). An unbalanced training set tends to overpredict the majority class for unknown sources (Tian et al. 2011).

To avoid this effect, we performed an oversampling of the AGN and stellar training sets so that in each considered magnitude bin we had a similar effective number of objects classified as galaxies, AGNs, and stars, respectively. In fact, despite our decision not to splits AGN and star classes into magnitude bins, unlike what we did in the case of galaxies, the imbalance between the numbers of representatives in each class remains high.

Using a simple oversampling technique, we raised the effective number of AGNs and stars up to ~80% of the number of galaxies in each magnitude bin considered. We therefore added in each magnitude bin a number of artificial objects calculated as

$$\lceil X_{\mathrm{i\_missing}} \rceil_{10} = NG_i \times 0.8 - X \quad (4)$$

where $X_{\mathrm{i\_missing}}$ is a number of missing objects (AGNs, stars), and symbol $\lceil \rceil_{10}$ corresponds to rounding the value up to the nearest ten. The additional artificial objects were created by shifting the observed magnitudes by an amount drawn from a Gaussian distribution with $\sigma = 0.05$. We also checked how the stellar and AGN training samples work if we did not perturb the colours, but instead populated real objects multiple times. As might be expected, the results of classifiers were worse than with randomly modified stars and AGNs. This method also allows us to take all possible small residuals differences into account in photometry between the two fields. Table 2 summarises the numbers of training galaxies in each magnitude-binned set together with the number of AGNs and stars after oversampling.

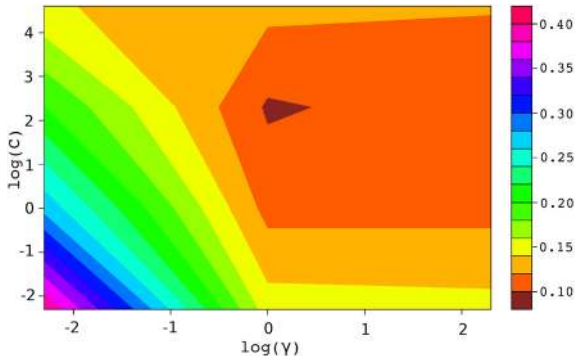

Fig. 8. Mean misclassification rate as a function of C and γ as estimated from the ten-fold cross-validation technique performed for each pair of parameters (see text for more details). The lower the ratio of misclassification, the better the performance of the SVM algorithm.

5. Results

5.1. Training procedure

To build a classifier that will be able to separate different classes of objects, it is necessary to tune the C and γ parameters using the training sample. For the best performance, we performed a grid search with values from $\gamma \in 10^{(-3:-1)}$ and $C \in 10^{(0:3)}$ using a ten-fold cross-validation technique. We first divided the full training sample into ten subsets of equal size and selected nine subsets to train the classification model and test it against the remaining subset (the so-called self-check). This test was repeated ten times, with a different subset removed for each training run. The classification accuracy was then averaged over the ten runs. This process was repeated for each value of the parameters C and γ. In Fig. 8 we present a representative plot of the the grid search, done for the apparent magnitude bin $19 \leqslant i' < 20$. The colour of each pointing of the grid codes the mean misclassification rate of all γ and C values (on a log scale on the X and Y axis, respectively). The misclassification rate is defined as (1-total accuracy) for each magnitude bin (see Eq. (6) further in the paper): the lower the ratio of misclassification, the better the performance of SVM algorithm. We would like to stress that a change in the parameter space (such as adding more parameters describing properties of sources) or a sufficient change in the number of training objects inside one class may result in altering the occupancy of training objects and therefore requires recalculating the best parameters.

To check the efficiency of our classifiers, we counted the true objects (true galaxies – TG, true AGNs – TAGN, and true stars – TS from the training sample originally classified as galaxies, AGNs, and stars, respectively) and false objects: FG (false

Table 3. Results of the self-check of the purely optical classifier (u^*, g', r', and i' only).

	$19 \leqslant i' < 20$			$20 \leqslant i' < 21$			$21 \leqslant i' < 22$			$22 \leqslant i' < 22.5$		
Total accuracy	85.01%			87.38%			85.09%			88.09%		
SVM/true	Galaxy	AGN	Star	Galaxy	AGN	Star	Galaxy	AGN	Star	Galaxy	AGN	Star
Number of sources	1884	1520	2 232	5483	4440	4440	6778	5440	5440	2126	1760	2232
Galaxy	**88.82**	15.70	10.98	**92.10**	6.23	15.06	**88.39**	15.50	10.01	**93.18**	17.47	3.00
AGN	4.45	**69.45**	10.23	3.28	**90.88**	4.48	4.04	**81.54**	3.81	4.37	**79.06**	3.28
Star	6.73	14.85	**78.79**	4.62	2.89	**80.46**	7.57	2.96	**86.19**	2.46	3.47	**93.72**

Notes. Columns corresponds to the true (spectroscopically classified) galaxies, stars, and AGNs. Rows correspond to objects classified as galaxies, AGNs, and stars by our classifier. Then values in bold correspond to the correctly classified objects (galaxies, AGNs, and stars) in defined i'-based apparent magnitude bins. Ratios of classified objects are given in percentage.

Table 4. Results of the self-check of the classifier with the NIR data (u^*, g', r', i', z', and K_s).

	$19 \leqslant i' < 20$			$20 \leqslant i' < 21$			$21 \leqslant i' < 22$			$22 \leqslant i' < 22.5$		
Total accuracy	95.47%			95.83%			94.28%			94.58%		
SVM/true	Galaxy	AGN	Star	Galaxy	AGN	Star	Galaxy	AGN	Star	Galaxy	AGN	Star
Number of sources	1884	1520	2232	5483	4440	4440	6778	5440	5440	2126	1760	2232
Galaxy	**96.28**	2.90	1.27	**97.61**	1.95	0.44	**97.11**	5.00	2.10	**96.10**	6.09	1.57
AGN	2.44	**95.91**	1.70	1.95	**96.34**	0.80	2.52	**94.83**	0.77	3.38	**92.94**	1.30
Star	1.28	1.19	**96.37**	0.44	0.27	**97.25**	0.37	0.17	**97.13**	0.52	0.97	**97.13**

Notes. Columns correspond to the true (spectroscopically classified) galaxies, stars, and AGNs. Rows correspond to objects classified as galaxies, AGNs, and stars by our classifier. The values marked in bold are correctly classified objects (galaxies, AGNs, and stars) in defined i'-based apparent magnitude bins. Ratio of classified objects are given in percentage.

galaxy: when a source from the stellar or AGN training sample is classified as a galaxy by the SVM); FS (false star: when an object from a galaxy or AGN training sample is classified as a star by the SVM); and FAGN (false AGN: when an object from a galaxy or star training sample is classified as an AGN by the SVM). We then calculated the accuracy of our classifier based on the formula:

$$\text{Accuracy} = \frac{\text{TG} + \text{TAGN} + \text{TS}}{\text{TG} + \text{TAGN} + \text{TS} + \text{FG} + \text{FAGN} + \text{FS}}. \quad (5)$$

After completing the ten-fold cross-validation process we calculated the total accuracy of the SVM classifier, defined as the mean accuracy for all iterations:

$$\text{Total Accuracy} = \frac{\Sigma_{i=1}^{N} \text{Accuracy}_i}{N}, \quad (6)$$

where $N = 10$ is the number of validation iterations. We performed this check in each magnitude bin considered.

In our work for galaxy/AGN/star classification, we used both a three- and five-dimensional colour space. The first one was built using only optical data, corresponding to $(u^* - g')$, $(g' - r)$, and $(r' - i)$ colours, while the second one included NIR data and thus used two extra colours: $(i' - z')$, and $(z' - K_s)$.

5.2. Optical $u^* g' r' i'$ classifier

We constructed colour–colour training samples without NIR data, based only on the optical u^*, g', r', and i' filter bands (a three-dimensional hyperspace). We found that the Total Accuracy, as well as the number of correctly classified objects for this approach, depend on the apparent magnitude of objects. Averaging over all magnitude bins ($19 \leqslant i' < 22.5$), once we average results by the number of objects in each bin, the mean Total Accuracy for the optical classifier is equal to 86.39%.

The results of the self-check of our classifier are shown in Table 3, showing that only in a few percent of the cases (less than 11% in all magnitude bins), galaxies are classified as a star or as an AGN. The most frequent misclassifications occur in the $19 \leqslant i' < 20$ bin, in which galaxies are correctly classified at the level of 88.82%, AGNs – 69.45%, and stars at the level of 78.79%. The misclassifications between stars and galaxies are noticeable in the first three bins. For $20 \leqslant i' < 21$ and $21 \leqslant i' < 22$ bins, more than 10% of spectroscopically classified stars are classified by the SVM as false galaxies (15.06% and 10.01%, respectively). In the same bins, AGNs are misclassified as galaxies at the high levels of 6.23% and 15.50%, respectively.

The misclassification of galaxies and AGNs happens mainly in the bins where the percentage of oversampled objects increases. The reason may be related either to our oversampling method or to the lower accuracy of photometry for the fainter sources, as well as to the intrinsic properties of classified sources in these bins. We stress that for the SVM method the 100% level of self-check is not desirable since it may indicate overfitting. The boundaries between different classes of objects defined by the training sample may become too rigid and artificially complex, not allowing for effective classification of real sources. Nevertheless, it seems that the present, very basic classifier, which was created on the basis similar to the standard colour–colour approach, works well for our training sample.

We next apply our trained classifier to VIPERS galaxies with redshift quality flag $\text{VIPERS}_{\text{Zflag}} = 3$, corresponding to a confidence of the redshift measurements – and correspondingly of correct identification as a galaxy – of >99% (hereafter GAL_3). Table 5 shows that GAL_3 are correctly classified at a level higher than 85% with a percentage of misclassification that is almost constant at a level of 15% maximum. The strong contamination by false stars is visible for objects fainter than $i' = 21$ mag. It is reassuring that this trend is similar to the self-check results (Table 3) demonstrating that the training sample is representative of the data. In the fainter magnitude bins, the photometric

Table 5. Test of SVM optical classifier on the galaxies with $\text{VIPERS}_{\text{Zflag}}$ equal to 3.

	$19 \leqslant i' < 20$	$20 \leqslant i' < 21$	$21 \leqslant i' < 22$	$22 \leqslant i' < 22.5$
Galaxies	90.97	91.41	85.38	88.82
False AGNs	2.76	2.81	3.06	4.45
False stars	6.27	5.78	11.56	6.73

Notes. In the first row we show the percentage of correctly classified galaxies. Second and third rows show the percentage of miss-classified galaxies: when a true galaxy is classified by SVM as an AGN or a star, respectively.

Table 6. Test of SVM classifier with NIR data on the galaxies with $\text{VIPERS}_{\text{Zflag}}$ equal to 3.

	$19 \leqslant i' < 20$	$20 \leqslant i' < 21$	$21 \leqslant i' < 22$	$22 \leqslant i' < 22.5$
Galaxies	95.38	95.17	93.09	92.72
False AGNs	2.42	2.72	4.30	5.29
False stars	2.20	2.11	2.61	1.99

Notes. The first row represents the percentage of correctly classified galaxies. Second and third rows show the percentage of mis-classified galaxies: when a true galaxy is classified by SVM as an AGN or a star, respectively.

errors increase such that the optical u^*, g', r', and i' fluxes are not as efficient in distinguishing galaxies and stars.

5.3. Optical+NIR ($u^* g' r' i' z' K_s$) classifier

We enlarged the parameter space by adding the NIR colours (z' and K_s) to our classifier (a five-dimensional hyperspace). We performed the same tests as for the optical classifier (self-check, and test on VIPERS GAL_3).

Our training sample, composed of exactly the same sources as the optical classifier, but with NIR measurements, allows us to train a new optical + NIR classifier. The mean Total Accuracy for this classifier is equal to 94.29%, i.e. higher than the pure optical one. Total accuracy for particular magnitude bins stays on the similar level ~95% for the whole i'-apparent magnitude binned sample. The constancy of the new classifier for objects fainter than 20 mag in i' band is very promising for the next tests and final classification of VIPERS objects.

Table 4 shows the self-check for the u^*, g', r', i', z' and K_s space classifier. When we average over all magnitude bins, galaxies are correctly classified in ~97.03%, AGNs in 95.13%, and stars in 97.05% of the cases. All these numbers are significantly higher than those for a purely optical classifier. In the case of AGNs, the difference between correctly classified sources for optical and optical+NIR classifiers is equal to 26.46%, 5.46%, 13.30%, and 13.88% for $19 \leqslant i' < 20$, $20 \leqslant i' < 21$, $21 \leqslant i' < 22$, and $22 \leqslant i' < 22.5$ apparent magnitude bins, respectively. Stars are correctly classified at a higher level than AGNs, with a difference between optical and optical+NIR classifiers equal to 17.58%, 16.79%, 10.94%, and 3.41% for the same magnitude bins.

Applying this classifier to VIPERS galaxies with $\text{VIPERS}_{\text{Zflag}}$ equal to 3 (GAL_3, Table 6) shows that galaxies are correctly classified at the very high level of 93.60% (we average results by the number of objects in each bin). Incorrect galaxy classifications, false AGNs and false stars, are very rare and do not exceed 2.65% for stars and 5.30% for AGNs.

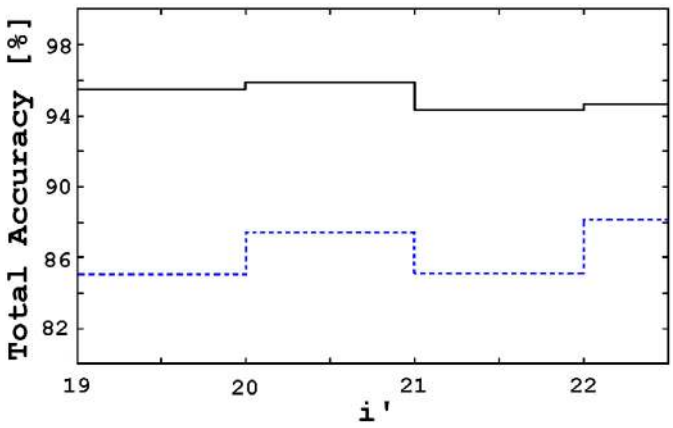

Fig. 9. Total accuracy for optical and optical+NIR classifiers (see Tables 3 and 4). Results for the optical classifier based on the u^*, g', r', and i' filter are marked as a dotted line. Solid line corresponds to the total accuracy of the optical+NIR classifier.

We can observe the trend for galaxies to have an increased risk of being misclassified as AGNs in the faintest magnitude bins. One possible explanation for this behaviour is the decrease in the quality of the photometry for the less luminous sources, which have a lower signal-to-noise ratio. On the other hand, the limiting magnitude of CFHTLS is much deeper than the VIPERS one, and photometry should still be fairly good down to mag i' 22.5. Another explanation could be that some of these galaxies are hosting faint AGNs that were not recognised during the visual verification and validation of the measured redshift, since with the decreasing luminosity the host galaxy becomes dimmer and the AGN component becomes more significant. This possibility will be examined further in future works.

5.4. Comparison of the classifiers

In Fig. 9 we compare the total accuracy for the optical and optical+NIR classifiers. However, on average the classifier based on the u^*, g', r', i', z', and K_s bands is 7.90% better then the classifier trained without z' and K_s data. Moreover, the total accuracy of the optical+NIR classifier decreases very weakly with the apparent magnitude, while a strong variation from bin-to-bin is visible for the purely optical classifier. etween the first and the second apparent magnitude bin the difference between their total accuracy rises from 6.49% to 10.46% from the fainter to the brighter bins.

The preponderance of the classifier constructed with the NIR data is confirmed by the efficiency of correctly classifying galaxies with $\text{VIPERS}_{\text{Zflag}}$ equal to 3 (GAL_3). Figure 10 shows the comparison of accuracy of both classifiers (with and without NIR data) for the GAL_3 sample. For the fainter objects ($21 \leqslant i' < 22$), the efficiency decreases rapidly for the classifier trained without z' and K_s bands, and much smoother for the more sophisticated classifier trained with infrared features.

We conclude that including NIR data to train the SVM algorithm significantly improves the efficiency of the galaxy/AGN/star classifier. It is evident that NIR features are very important for building an effective classifier for basic astronomical classification of these three classes of sources. Based on the above tests, we decided to choose the classifier based on the u^*, g', r', i', z', and K_s bands to be used in our next analysis.

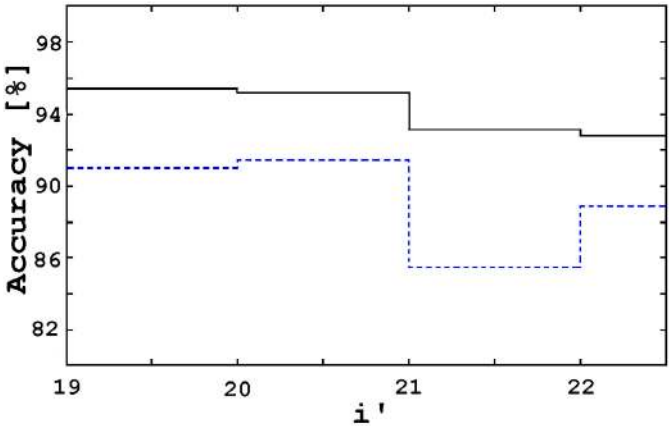

Fig. 10. Accuracy of optical and optical+NIR classifiers for VIPERS galaxies with $\mathrm{VIPERS_{Zflag}}$ equal to 3 ($\mathrm{GAL_3}$). Results for classifier based on the u^*, g', r', and i' filters only are marked as a dotted line. Solid line corresponds to the classifier with the NIR data (u^*, g', r', i', z', and K_s).

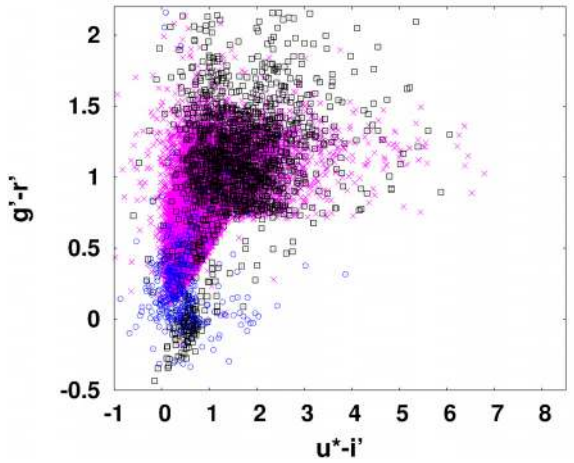

Fig. 11. Representative colour–colour plot for all objects used for a consistency check for VIPERS objects with redshift confirmation levels >99%, with i' apparent magnitude between 19 and 22.5. Pink x-s represents galaxies with $\mathrm{VIPERS_{Zflag}} = 3$. Open blue circles correspond to AGN sample with redshift confirmation level equal to or higher than 99% ($\mathrm{VIPERS_{Zflag}}$ equal to 13 and 14). Open black squares correspond to stellar sample with $\mathrm{VIPERS_{Zflag}}$ equal to 3 and 4.

6. Consistency checks on VIPERS data

6.1. VIPERS objects with redshift confirmation level of ≧99%

We now apply the optical+NIR classifier only to VIPERS data:

- galaxy sample – all ($\mathrm{GAL_3}$) galaxies in i'-apparent magnitude range between 19 and 22.5 mag, with the total number of sources equal to 13 539,
- AGN sample – all AGNs detected by VIPERS, with redshift confirmation level equal to or higher than 99%, and with i' apparent magnitude between 19 and 22.5 (367 objects). All of these AGNs were used to build the training sample (see Sect. 4.2) which means that our classifier should know their position in our five-dimensional space of parameters. This is not as worrisome as it may look thanks to the high oversampling needed for AGN sample (more than 200% for the brightest and the faintest apparent magnitude bins, and almost 800% for $20 \leqslant i' < 21$ and $21 \leqslant i' < 22$ for i'-apparent magnitude bins) that significantly erases the possibly peculiar characteristics of the 367 AGN chosen for the training sample.
- stellar sample – all spectroscopically detected stars, with confirmation level of >99% ($\mathrm{VIPERS_{Zflag}}$ equal to 3 and 4), and i' apparent magnitude between 19 and 22.5 (1729 stars). All stars with $\mathrm{VIPERS_{Zflag}} = 4$ were used as a part of stellar training sample.

Figure 11 shows the representative colour–colour plot for $\mathrm{GAL_3}$, AGNs with $\mathrm{VIPERS_{Zflag}}$ equal to 13 and 14, and stars with $\mathrm{VIPERS_{Zflag}}$ equal to 3 and 4, chosen for the consistently check.

For this test, all three classes of sources were divided into four i'-apparent magnitude bins ($19 \leqslant i' < 20$, $20 \leqslant i' < 21$, $21 \leqslant i' < 22$, and $22 \leqslant i' < 22.5$), the same as used in the training sample. Then, we applied our optical+NIR classifier to this data. Table 7 shows the results of the automatic classification.

The mean accuracy for galaxies, averaged over the mean number of objects in each apparent magnitude bin, equals 93.60%. This result for galaxy classification displays only aslightly lower level of efficiency (~1.50%) than the galaxy classification obtained during the self-check of the classifier (see Sect. 5.3). It means that the hyperspace of galaxy parameters used for the training sample is well defined.

The result of AGN classification is worse than the one obtained during the self-check but still satisfactory. After averaging over all magnitude bins, AGNs are correctly classified at a level equal to 81.80% with a significant decrease with i' apparent magnitude between 21 and 22 mag. Stars are correctly classified at the high mean level of 92.52% with a significant drop for the $22 \leqslant i' < 22.5$ apparent magnitude bin (84.47%). The performance of the classifier in the case of AGNs may look relatively poor. However, as already mentioned, we should remember that the VIPERS selection allows AGNs preclassified as galaxies or stars based on their colour properties. Keeping this in mind, we should instead feel satisfied that a high fraction of these AGNs can be separated into a different section of the five-dimensional hyperspace from galaxies and stars, when using an AGN training sample that only consists of 498 objects.

We did not find any crucial misclassifications for the galaxy sample. The galaxies are classified correctly on a very high level. For the AGN sample, the contamination of true AGNs classified as galaxies (8.17%, 7.37%, 10.46%, 14.90% for the $19 \leqslant i' < 20$, $20 \leqslant i' < 21$, $21 \leqslant i' < 22$, and $22 \leqslant i' < 22.5$ bins, respectively) and stars (8.96%, 10.55%, 6.79%, 9.56% for the $19 \leqslant i' < 20$, $20 \leqslant i' < 21$, $21 \leqslant i' < 22$, and $22 \leqslant i' < 22.5$ bins, respectively) is significant. For the stellar sample, the classifier misclassified true stars as galaxies more often than AGNs. In the future development of this classifier, we will include the morphological information, as well as emission/absorption lines, which should improve the algorithm and increase the percentage of correctly classified sources as well. Including the morphological information will allow us to construct a classifier that could be applied to purely photometric surveys, similar to the one presented in this paper. Adding spectroscopic information to the parameter space would restrict the use of the classifier, but it would allow for more precise classification schemes.

6.2. VIPERS objects with redshift confirmation level lower than 99%

We performed a classification for VIPERS objects with confirmation levels lower than 99%. In particular, we used galaxies,

Table 7. Results of the test of the optical+NIR classifier for GAL_3, and AGNs and stars with redshifts measurements on a confirmation level ≥ to 99%.

	$19 \leqslant i' < 20$			$20 \leqslant i' < 21$			$21 \leqslant i' < 22$			$22 \leqslant i' < 22.5$		
SVM/true	Galaxy	AGN	Star	Galaxy	AGN	Star	Galaxy	AGN	Star	Galaxy	AGN	Star
Number of sources	445	69	337	3 271	1340	428	7 667	127	701	2 156	37	263
Galaxy	**95.38**	12.52	4.17	**95.17**	7.37	3.27	**93.09**	10.46	3.42	**92.72**	14.90	9.09
AGN	2.42	**77.34**	3.70	2.72	**82.08**	3.27	4.30	**82.75**	1.43	5.29	**75.54**	6.44
Star	2.20	10.14	**92.13**	2.11	10.55	**93.46**	2.61	6.79	**95.15**	1.99	9.56	**84.47**

Notes. Values marked in bold correspond to the correctly classified objects (galaxies, AGNs, and stars) in i'-based apparent magnitude bins. The ratio of the classified objects is given in percentage.

Table 8. Results of the optical + NIR classifier for galaxies, AGNs, and stars with redshifts measurements on a confirmation level equal to 95% ($VIPERS_{Zflag} = 2$).

	$19 \leqslant i' < 20$			$20 \leqslant i' < 21$			$21 \leqslant i' < 22$			$22 \leqslant i' < 22.5$		
SVM/true	Galaxy	AGN	Star	Galaxy	AGN	Star	Galaxy	AGN	Star	Galaxy	AGN	Star
Number of sources	8	33	10	945	75	27	5757	80	145	6226	48	159
Galaxy	**84.15**	6.25	30.00	**94.18**	12.00	22.22	**92.81**	20.25	29.65	**58.62**	4.17	17.61
AGN	9.75	**93.75**	40.00	3.49	**88.00**	29.63	4.41	**77.22**	11.03	20.56	**93.75**	18.87
Star	6.10	0.00	**30.00**	2.33	0.00	**48.15**	2.78	2.53	**59.32**	20.82	2.08	**63.52**

Notes. Objects are not related to the training sample. Values marked in bold correspond to the correctly classified objects (galaxies, AGNs, and stars) in i'-based apparent magnitude bins. The ratio of the classified objects is given in percentage.

Table 9. Results of the optical + NIR classifier for galaxies, AGNs, and stars with redshifts measurements on a confirmation level equal to 50% ($VIPERS_{Zflag}$ equals to 1).

	$19 \leqslant i' < 20$			$20 \leqslant i' < 21$			$21 \leqslant i' < 22$			$22 \leqslant i' < 22.5$		
SVM/true	Galaxy	AGN	Star	Galaxy	AGN	Star	Galaxy	AGN	Star	Galaxy	AGN	Star
Number of sources	35	8	4	355	13	24	2833	35	81	3157	30	139
Galaxy	**85.71**	50.00	25.00	**92.68**	23.08	37.50	**88.14**	20.00	60.56	**43.09**	23.33	25.90
AGN	11.43	**50.00**	75.00	3.94	**76.92**	37.50	7.12	**68.57**	9.86	31.03	**66.67**	31.65
Star	2.86	0.00	**0.00**	3.38	0.00	**25.00**	4.74	11.43	**29.58**	25.88	10.00	**42.45**

Notes. Objects are not connected with the training sample. Values marked in bold correspond to the correctly classified objects (galaxies, AGNs, and stars) in i'-based apparent magnitude bins. The ratio of the classified objects is given in percentage.

AGNs, and stars from the VIPERS database, with the quality of the measured redshift, $VIPERS_{Zflag}$, equal to two and one. $VIPERS_{Zflag}$ equals two means that the measured redshift is fairly secure, with a confidence level ⩾95%. Objects with $VIPERS_{Zflag}$ equal to one are more tentative, and their redshift measurement was based on weak spectral features and/or continuum shape. For these objects there is a ∼50% probability that the redshift could be wrong. A more detailed description of the $VIPERS_{Zflag}$ and quality of measured redshifts can be found in Guzzo et al. (2013), and Garilli et al. (2012).

Results of SVM classification of objects with $VIPERS_{Zflag}$ equal to two (Table 8) and one (Table 9) show very good conformity to the previously user supervised estimations. Galaxies are classified with agreement to redshift measurements on the mean level of 76.45% for $VIPERS_{Zflag} = 2$ and 66.08% for $VIPERS_{Zflag} = 1$.

The ongoing scientific analysis of galaxy evolution and clustering is mainly based on objects that have secure redshift measurements ($VIPERS_{Zflag} \geq 2$, depending on the topic). With the SVM classification, we can reconfirm the identify of galaxies with the lower quality flags and thus increase the number of galaxies that could be used for more detailed analysis. This may apply to 4735 galaxies, 58 AGNs, and 86 stars with $VIPERS_{Zflag}$ equal to one[8]. This method may also reconfirm the class of 9952 galaxies, 177 AGNs, and 160 stars with $VIPERS_{Zflag} = 2$ classified as galaxies, AGNs, and stars by checking the results twice by different observers, and by our classifier.

One may argue that the $VIPERS_{Zflag}$ is related to the redshift value, not to the identification of the galaxy itself. However, it should be noted that a majority of sources with low $VIPERS_{Zflag}$ are absorption line systems with noisy, low signal-to-noise spectra. Galaxies with such spectra can be particularly easily misclassified as stars during the spectroscopic measurement process, either automatic or human-supervised. For instance, typical features of an elliptical galaxy at $z \sim 1$, around the Balmer break, can be confused with characteristic features of an M-type star. To confirm the redshift measurements or flag validation, it is possible to use SED templates for the photometric redshift estimation, and to compare spectroscopic redshifts with photometric ones, but SED-fitting for sources with poor photometry can be degenerate possibly leading to biased results. In such cases, an independent confirmation that the position of an object in

[8] These numbers were calculated as a sum of galaxies, AGNs, and stars which were classified to the same class of objects during redshift validation and by an optical+NIR classifier; marked in bold in Table 9.

the five-dimensional colour space is actually typical of galaxy, and actually increases also probability that its redshift has been assigned correctly.

The number of stars and AGNs in the sample of objects with $\mathrm{VIPERS}_{\mathrm{Zflag}}$ equal to one and two is very low (a few objects in the brightest apparent luminosity bin, and a few dozen for objects with i' magnitude lower than 21 mag). This fact results from an initial star/galaxy separation performed by VIPERS. In the VIPERS database, the stars, which remained after the colour–colour preselection, are not typical, and they occupy a similar area to galaxies on the colour–colour plots. Then, that we can reconfirm the identity of a significant fraction of them can already be regarded as a success. The spectra of stars with $\mathrm{VIPERS}_{\mathrm{Zflag}} = 1$ or 2, which were classified as galaxies by our classifier, will have to be re-examined since some of them might also be genuine galaxies.

We should consider that the VIPERS galaxy sample is not pure and includes some AGN types such as those with narrow-line features, even for objects with $\mathrm{VIPERS}_{\mathrm{Zflag}} > 3$. During the standard redshift measurements process only the broad line AGNs are being recognised and flagged. This implies that that our galaxy training sample alco contains, in addition to a pure sample of normal galaxies, specific types of AGNs, otherwise difficult to recognise in VIPERS spectra during a standard redshift measurements process. The VIPERS hunt for the specific types of AGNs lurking within the heap of collected sources is still going on, so we are forced to work with the data composed of both galaxies and AGNs, at least for now. The contamination of galaxies and AGNs is most prominent for the faintest bin ($22 \leqslant i' < 22.5$), where more than 20% (30%) of objects classified as galaxies with $\mathrm{VIPERS}_{\mathrm{Zflag}}$ equal to two (one) are identified as AGNs by an optical+NIR SVM classifier.

We look forward to using SVM methods to add more information on spectral lines and source morphologies as a very promising tool to improve classification for fainter sources and to refine further classes of objects that the software can discriminate.

7. Comparison with combined spectral energy distribution fitting and geometric method

As a test of efficiency of the SVM VIPERS classifier, we compared our algorithm with the star/galaxy separtion of the VVDS data performed by Coupon (Coupon et al. 2009; Guzzo et al. 2013). We computed the incompleteness of our galaxy selection as the ratio of true galaxies/AGNs lost after SVM classification, and we defined contamination as a number of stars mis-classified by SVM as galaxies/AGNs.

Coupon et al. (2009) base their star/galaxy classification on the most secure spectroscopic sample from the VVDS F02 and VVDS F22 fields. The method adopted for the star/galaxy separation was a combination of a geometric method for objects brighter than $i' = 21$ mag (half-light radius parameter, r_{h}, defined as the radius containing half of the object's flux, which was provided by the CFHTLS database), and a combination of geometric and photometric methods for objects fainter than $i' = 21$ mag, fitting u^*, g', r', i', and z' bands by a set of the SED templates with the Le Phare photometric redshift code (Arnouts et al. 1999; Ilbert et al. 2006). For a detailed description of this method we refer the reader to Guzzo et al. (2013).

7.1. Sample selection

We performed a star/non-star (where non-star for our classifier means galaxy+AGN) selection using the VVDS Deep F02 survey matched with the CFHTLS photometric catalogue (T0005 data release). We decided to perform star/galaxy classification in the VVDS Deep F02 only, because the stellar training sample used for our classifier was built from stars from the VVDS Wide F22 field. Only a part of AGNs from the VIPERS survey was used to train our algorithm. As a result, our stellar/galaxy separation in this field would be treated preferentially, which could bias the results.

For our test we selected objects with the most secure VVDS flags ($\mathrm{VVDS}_{\mathrm{Zflag}}$ equal to 3, and 4). In the next step we selected objects using the same colour/redshift criteria as applied to the VIPERS survey:

$$(r - i) > 0.5 \times (u - g) \text{ or } (r - i) > 0.7. \tag{7}$$

For the more detailed description and the origin of this colour-based selection, we refer the reader to Guzzo et al. (2013). Then, we divided our sample into two subsamples:

1. non-stars with spectroscopic redshift $\geqslant$0.01; and
2. stars with spectroscopic redshifts $\leqslant$0.01;

and then into i' apparent magnitude-binned samples. We stress that the ratio of AGNs is not known within the galaxy sample in this case.

7.2. Method

The SVM opt+NIR galaxy/AGN/star classifier was applied to this data set. We computed the incompleteness and contamination for selected non-stars (galaxies and AGNs). Since we did not use geometric selection based on the r_{h}, we decided to perform the comparison with the Coupon et al. (2009) method for only the fainter bins ($21 \leqslant i' < 22$, and $22 \leqslant i' < 22.5$), where star/galaxy separation was performed based not only on the geometrical properties of sources, but also by fitting SEDs. We defined the incompleteness (INC) and contamination (CON) ratio, following Coupon et al. (2009) and Guzzo et al. (2013), as

$$\mathrm{INC} = \frac{\mathrm{NG}_{\mathrm{true}} - \mathrm{NG}_{\mathrm{SVMgood}}}{\mathrm{NG}_{\mathrm{true}}}, \tag{8}$$

and

$$\mathrm{CON} = \frac{\mathrm{NG}_{\mathrm{SVMbad}}}{\mathrm{NG}_{\mathrm{SVMestimated}}}, \tag{9}$$

where

- $\mathrm{NG}_{\mathrm{true}}$ is a total number of spectroscopically classified non-stars in the VVDS Deep F02 field with the most secure redshift quality flag ($\mathrm{VVDS}_{\mathrm{Zflag}}$3 and 4),
- $\mathrm{NG}_{\mathrm{SVMgood}}$ – a number of real non-stars (galaxies and AGNs) classified by SVM algorithm as non-stellar objects,
- $\mathrm{NG}_{\mathrm{SVMbad}}$ – a number of galaxies/AGNs mis-classified by our classifier as stars, and
- $\mathrm{NG}_{\mathrm{SVMestimated}}$ is a total number of objects classified by SVM as a galaxies or AGNs.

Table 10. (Galaxy+AGN)/star selection results: the incompleteness and contamination of the VIPERS galaxy sample (VVDS-Deep F02 field) expected from the star-galaxy separation process adopted in VIPERS, and from the SVM opt+NIR classifier.

Apparent magnitude	INC	CON	INC	CON
	Guzzo et al. (2013)		SVM	
$21 \leqslant i' < 22$	2.07%	0.87%	2.13%	2.39%
$22 \leqslant i' < 22.5$			2.05%	2.00%

7.3. Results

Table 10 shows the results of incompleteness and contamination of two classifiers applied to the VVDS-Deep F02 sample: one based on the SEDs and geometrical properties of sources (Guzzo et al. 2013) and the other based on the SVM method. Presented values are the expected once for incompleteness and contamination of the star/galaxy separation in the W1 VIPERS field.

The incompleteness for both methods is similar (2.07% from Guzzo et al. 2013 vs. 2.13% and 2.05 for SVM). The stellar contamination in the SVM method is slightly higher (0.87% vs. 2.39% and 2.00% for SVM algorithm). Comparing these results with the results of the self-check of our classifier (see Table 4), we conclude that the high stellar contamination for these bins might be related to misclassifications between AGNs and stars[9]. Unfortunately, this conclusion cannot be compared directly with the Coupon et al. (2009) method because theirs does not classify AGNs.

We checked the real stellar contamination in the W1 field after VIPERS spectroscopic measurements. In total, 264 from the 23 360 objects in the $21 \leqslant i' < 22.5$ apparent magnitude bin preclassified as galaxies using the SEDs+r_h method from the PDR-1 catalogue were spectroscopically classified as a stars with $\mathrm{VIPERS}_{\mathrm{Zflag}} \geqslant 1$. It means that the real contamination on W1 VIPERS filed for object fainter than $i' \geqslant 21$ mag is equal to

$$\mathrm{CON}_{\mathrm{VIPERS\_W1\_i'}} \geqslant 21 = \frac{264}{23\,360} = 1.13. \quad (10)$$

This value is between the contamination factor calculated from Guzzo et al. (2013) and the one given by our SVM opt+NIR classifier. Taking this difference into account, we conclude that the results obtained by both methods are similar and very close to the real values obtained from the spectroscopic observations.

We performed a classification of 264 objects preclassified as galaxies through the SEDs+r_h method, and spectroscopicaly classified as a stars. In total 122 sources from this sample (46.6%) were correctly classified as stars by our algorithm. Taking only sources with very high confidence level of spectroscopic classification into account ($\mathrm{VIPERS}_{\mathrm{Zflag}} \geqslant 3$; 123 sources), we found that our algorithm shows 74.8% of accuracy in correctly classifying 92 of those objects as stars.

It confirms that our SVM classification, based on spectroscopically measured objects from the VIPERS and VVDS surveys, can provide an efficient star/no-star classifier. This method is also very fast. The only time-consuming part of the SVM-based method is the tuning of the classifier, but once the classifier is trained, all the following classifications are very fast and can be done without any additional supervision.

[9] More than 5% of real AGNs were classified by our algorithm as stars.

8. Conclusions

Application of the SVM algorithm can deliver an excellent (with accuracy level for self-check test higher that 98% for galaxies, 94% for AGNs, and 93% for stars) classification for three classes of objects, after a careful selection of the training sample. For our analysis we constructed two classifiers, with and without near infrared data using a multidimensional colour hyperspace. A part of the AGN and star samples were extracted from the VVDS survey. We have found a significant improvement in the SVM classification (8% in the total accuracy of the classifier) adding an NIR colour parameters to our feature vectors.

For the optical+NIR classifier, we obtained very good agreement (93.60%, 81.80%, and 92.52% for galaxies, AGNs, and stars, respectively) with the VIPERS spectroscopic sample with flag confidence level of z measurements equal to 95%. What makes our approach to SVM classification more suitable is that the enormous amount of excellent quality data, means that we could create the classifier, which was trained on the part of the most secure sources, and then test it against the remaining secure objects to create the most efficient pattern recognition system. The VIPERS survey gathered a large number of sources (55 358) with very good spectroscopic measurements, which then were strictly analysed to obtain the most secure redshifts. This allowed for the choice of the best sample, which could be used as a basis for the new methods of automatic classification.

SVM classifiers are mostly used in the literature for separating two classes of sources (e.g. stars and galaxies). The only recent application of the SVM to the galaxy/AGNs/stars classification was performed by Saglia et al. (2012), who trained and used his classifier for the Pan-STARRS1 data. Comparing the accuracies of our classifier and those of Saglia et al. (2012) we found that our self-check results look somewhat better (97%, 95%, 97% vs 97%, 84%, 85% for galaxies, AGNs, stars for VIPERS and Pan-STARSS1 classifier, respectively). However, we have to stress that both methods cannot be directly compared because of initial differences in both surveys. Pan-STARRS1 is a magnitude-limited survey, which implies a much higher variety of properties of all the sources it contains. In contrast, VIPERS was preselected to contain only $0.5 < z < 1.2$ galaxies, witch assures that they form a much more distinct and better separated group in a multicolour space. This may facilitate a separation between galaxies and AGNs, as well as a part of stars that were re-introduced to the VIPERS target sample as AGN candidates. On the other hand, the lack of "typical" stars in the VIPERS database (rejected after colour and half-light radius preselection) occupying the same colour–colour space as galaxies may hamper our classification based only on colours, and decrease the efficiency of our classifier for sources from the real sample. The difference in the performance with respect to the PAN-STARS1 SVM method might also be related to the different broad-band photometry. The tests of accuracy of our purely optical ($u^*g'r'i'$) classifier show similar efficiency to the PAN-STARS1 results (94%, 82%, and 93% for galaxies, AGNs, and stars from VIPERS survey), while the dimension of PAN-STARS1 parameter space is higher than ours (4D in case of PAN-STARS1 and 3D in the case of VIPERS optical classifier). It suggests that the key points of our method might be a more suitable photometry ($u*$ instead of z_{P1} and y_{P1} bands) and division of our sample into apparent magnitude bins.

Our approach allows us to photometrically classify sources in the VIPERS survey, augmenting the spectral information. By classifying the sources with low-quality spectra, we can improve the classification and enlarge the samples that may be

used for analysis. Using the optical+NIR classifier, we confirmed the class of 4900 objects with low flags. Further improvement in our classifier by the addition of the morphology and emission/absorption line information will improve the already very good performance of galaxy/AGN/star classifier. It will also allow for developing a more specific galaxy and AGN-type classifications.

Acknowledgements. We acknowledge the crucial contribution of the ESO staff for the management of service observations. In particular, we are deeply grateful to M. Hilker for his constant help and support of this programme. Italian participation in VIPERS has been funded by INAF through the PRIN 2008 and 2010 programmes. L.G. and B.R.G. acknowledge support by the European Research Council through the Darklight ERC Advanced Research Grant (# 291521). O.L.F. acknowledges the support of the European Research Council through the EARLY ERC Advanced Research Grant (# 268107). Polish participants have been supported by the Polish Ministry of Science (grant N N203 51 29 38), the Polish-Swiss Astro Project (co-financed by a grant from Switzerland, through the Swiss Contribution to the enlarged European Union), the European Associated Laboratory Astrophysics Poland-France HECOLS, and the Japan Society for the Promotion of Science (JSPS) Postdoctoral Fellowship for Foreign Researchers (KM, P11802). A.S. has been supported by the Global COE Program Request for Fundamental Principles in the Universe: from Particles to the Solar System and the Cosmos commissioned by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan, and by the JSPS Strategic Young Researcher Overseas Visits Program for Accelerating Brain Circulation, Construction of a Global Platform for the Study of Sustainable Humanosphere. G.D.L. acknowledges financial support from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no. 202781. W.J.P. and R.T. acknowledge financial support from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no. 202686. W.J.P. is also grateful for support from the UK Science and Technology Facilities Council through the grant ST/I001204/1. E.B., F.M., and L.M. acknowledge the support from grants ASI-INAF I/023/12/0 and PRIN MIUR 2010-2011. Y.M. acknowledges support from CNRS/INSU (Institut National des Sciences de l'Univers) and the Programme National Galaxies et Cosmologie (PNCG). C.M. is grateful for support from specific project funding by the Institut Universitaire de France and the LABEX OCEVU.

References

Akbani, R., Kwek, S., & Japkowicz, N. 2004, in Proceedings of the 15th European Conference on Machine Learning (ECML), 39
Arnouts, S., Cristiani, S., Moscardini, L., et al. 1999, MNRAS, 310, 540
Ball, N. M., Brunner, R. J., Myers, A. D., & Tcheng, D. 2006, ApJ, 650, 497
Beckwith, S. V. W., Stiavelli, M., Koekemoer, A. M., et al. 2006, AJ, 132, 1729
Bel, J., et al. 2013, A&A, submitted
Bland-Hawthorn, J. 2012, RA&A, 12, E1
Boulade, O., Charlot, X., Abbon, P., et al. 2000, in SPIE Conf. Ser. 4008, eds. M. Iye, & A. F. Moorwood, 657
Brightman, M., & Nandra, K. 2012, MNRAS, 422, 1166
Chang, C.-C., & Lin, C.-J. 2011, ACM Transactions on Intelligent Systems and Technology, 2, 27, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, issue = 3
Chiu, K., Zheng, W., Schneider, D. P., et al. 2005, AJ, 130, 13
Colless, M., Dalton, G., Maddox, S., et al. 2001, MNRAS, 328, 1039
Colless, M., Peterson, B. A., Jackson, C., et al. 2003 [arXiv:astro-ph/0306581]
Coupon, J., Ilbert, O., Kilbinger, M., et al. 2009, A&A, 500, 981
Cristianini, N., & Shawe-Taylor, J. 2000, An Introduction to Support Vector Machines: and Other Kernel-Based Learning Methods (Cambridge University Press)
Davidzon, I., Bolzonella, M., et al. 2013, A&A, in press, DOI: 10.1051/0004-6361/201321511
de la Torre, S., Guzzo, L., Peacock, J. A., et al. 2013, A&A, in press, DOI: 10.1051/0004-6361/201321463
Drinkwater, M. J., Gregg, M. D., Hilker, M., et al. 2003, Nature, 423, 519
Emerson, J., & Sutherland, W. 2010, The Messenger, 139, 2
Fadely, R., Hogg, D. W., & Willman, B. 2012, ApJ, 760, 15
Fritz, A., Scodeggio, M., et al. 2013, A&A, submitted
Garilli, B., Le Fèvre, O., Guzzo, L., et al. 2008, A&A, 486, 683
Garilli, B., Paioro, L., Scodeggio, M., et al. 2012, PASP, 124, 1232
Gavignaud, I., Bongiorno, A., Paltani, S., et al. 2007, VizieR Online Data Catalog, 345, 70079
Goranova, Y., Hudelot, P., Contini, T., et al. 2009, The CFHTLS T0006 Release, http://terapix.iap.fr/cplt/table_syn_T0006.html
Guzzo, L., Scodeggio, M., Garilli, B., et al. 2013, A&A, submitted [arXiv:1303.2623]
Hassan, T., Mirabal, N., Contreras, J. L., & Oya, I. 2013, MNRAS, 428, 220
Henrion, M., Mortlock, D. J., Hand, D. J., & Gandy, A. 2011, MNRAS, 412, 2286
Hsu, C.-W., Chang, C. C., & C.-J., L. 2010, A Practical Guide to Support Vector Classification, Department of Computer Science, National Taiwan University, Taiwan, http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf
Huertas-Company, M., Rouan, D., Tasca, L., Soucail, G., & Le Fèvre, O. 2008, A&A, 478, 971
Ilbert, O., Arnouts, S., McCracken, H. J., et al. 2006, A&A, 457, 841
Ivezic, Z., Tyson, J. A., Axelrod, T., et al. 2009, in BAAS, 41, Am. Astron. Soc. Meet. Abstracts, 213, 460.03
Kaiser, N., Burgett, W., Chambers, K., et al. 2010, in SPIE Conf. Ser., 7733
Kron, R. G. 1980, ApJS, 43, 305
Laureijs, R., Gondoin, P., Duvet, L., et al. 2012, in SPIE Conf. Ser., 8442
Le Fèvre, O., Saisse, M., Mancini, D., et al. 2000, in SPIE Conf. Ser. 4008, eds. M. Iye, & A. F. Moorwood, 546
Le Fèvre, O., Vettolani, G., Garilli, B., et al. 2005, A&A, 439, 845
Marchetti, A., Granett, B. R., Guzzo, L., et al. 2012, MNRAS, 107
Marulli, F., Bolzonella, M., Branchini, E., et al. 2013, A&A, 557, A17
Mellier, Y., Bertin, E., Hudelot, P., et al. 2008, The CFHTLS T0005 Release, http://terapix.iap.fr/cplt/oldSite/Descart/CFHTLS-T0005-Release.pdf
Meyer, D. 2001, R News, 1, 23
Mohr, J. J., Armstrong, R., Bertin, E., et al. 2012, in Software and Cyberinfrastructure for Astronomy II., SPIE Conf. Ser., 8451
Peng, N., Zhang, Y., Zhao, Y., & Wu, X.-B. 2012, MNRAS, 425, 2599
Pierre, M., Valtchanov, I., Altieri, B., et al. 2004, J. Cosmol. Astropart. Phys., 9, 11
Pollo, A., Rybka, P., & Takeuchi, T. T. 2010, A&A, 514, A3
Puget, P., Stadler, E., Doyon, R., et al. 2004, in SPIE Conf. Ser. 5492, eds. A. F. M. Moorwood, & M. Iye, 978
Raskutti, B., & Kowalczyk, A. 2004, SIGKDD Explor. Newsl., 6, 60
Richards, G. T., Fan, X., Newberg, H. J., et al. 2002, AJ, 123, 2945
Saglia, R. P., Tonry, J. L., Bender, R., et al. 2012, ApJ, 746, 128
Schlegel, D. J., Finkbeiner, D. P., & Davis, M. 1998, ApJ, 500, 525
Scodeggio, M., Franzetti, P., Garilli, B., Le Fèvre, O., & Guzzo, L. 2009, The Messenger, 135, 13
Shawe-Taylor, J., & Cristianini, N. 2004, Kernel Methods for Pattern Analysis (Cambridge University Press)
Sholl M. J., Ackerman M. R., Bebek C., et al. 2012, in Ground-based and Airborne Instrumentation for Astronomy IV, Proc. SPIE 8446, 844667
Solarz, A., Pollo, A., Takeuchi, T. T., et al. 2012, A&A, 541, A50
Stern, D., Eisenhardt, P., Gorjian, V., et al. 2005, ApJ, 631, 163
Stern, D., Assef, R. J., Benford, D. J., et al. 2012, ApJ, 753, 30
Tang, Y., Zhang, Y.-Q., Chawla, N. V., & Krasser, S. 2009, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 39, 281
Thibault, S., Cui, Q., Poirier, M., et al. 2003, in SPIE Conf. Ser. 4841, eds. M. Iye, & A. F. M. Moorwood, 932
Tian, J., Gu, H., & Liu, W. 2011, Neural Comput. Appl., 20, 203
Vapnik, V. N. 1995, The Nature of Statistical Learning Theory (Springer)
Vanschoenwinkel, B., & Manderick, B. 2005, in Proc. First international conference on Deterministic and Statistical Methods in Machine Learning, 256
Vasconcellos, E. C., de Carvalho, R. R., Gal, R. R., et al. 2011, AJ, 141, 189
Walker, H. J., Volk, K., Wainscoat, R. J., Schwartz, D. E., & Cohen, M. 1989, AJ, 98, 2163
Wittman, D. M., Tyson, J. A., Dell'Antonio, I. P., et al. 2002, in SPIE Conf. Ser. 4836, eds. J. A. Tyson, & S. Wolff, 73
Woźniak, P. R., Williams, S. J., Vestrand, W. T., & Gupta, V. 2004, AJ, 128, 2965

[1] Department of Particle and Astrophysical Science, Nagoya University, Furo-cho, Chikusa-ku, 464-8602 Nagoya, Japan
e-mail: malek.kasia@nagoya-u.jp
[2] Astronomical Observatory of the Jagiellonian University, Orla 171, 30-001 Cracow, Poland
[3] National Centre for Nuclear Research, ul. Hoza 69, 00-681 Warszawa, Poland
[4] INAF – Istituto di Astrofisica Spaziale e Fisica Cosmica Milano, via Bassini 15, 20133 Milano, Italy

[5] Aix Marseille Université, CNRS, LAM (Laboratoire d'Astrophysique de Marseille) UMR 7326, 13388 Marseille, France
[6] INAF – Osservatorio Astronomico di Brera, via Brera 28, 20122 Milano, via E. Bianchi 46, 23807 Merate, Italy
[7] INAF – Osservatorio Astrofisico di Torino, 10025 Pino Torinese, Italy
[8] Canada-France-Hawaii Telescope, 65–1238 Mamalahoa Highway, Kamuela, HI 96743, USA
[9] Aix-Marseille Université, CNRS, CPT (Centre de Physique Théorique) UMR 7332, 13288 Marseille, France
[10] INAF – Osservatorio Astronomico di Bologna, via Ranzani 1, 40127 Bologna, Italy
[11] Dipartimento di Matematica e Fisica, Università degli Studi Roma Tre, via della Vasca Navale 84, 00146 Roma, Italy
[12] INFN, Sezione di Roma Tre, via della Vasca Navale 84, 00146 Roma, Italy
[13] INAF – Osservatorio Astronomico di Roma, via Frascati 33, 00040 Monte Porzio Catone (RM), Italy
[14] Laboratoire Lagrange, UMR7293, Université de Nice Sophia-Antipolis, CNRS, Observatoire de la Côte d'Azur, 06300 Nice, France
[15] Institute of Astronomy and Astrophysics, Academia Sinica, PO Box 23-141, Taipei 10617, Taiwan
[16] Dipartimento di Fisica e Astronomia – Università di Bologna, viale Berti Pichat 6/2, 40127 Bologna, Italy
[17] INAF – Osservatorio Astronomico di Trieste, via G. B. Tiepolo 11, 34143 Trieste, Italy
[18] SUPA, Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK
[19] Dipartimento di Fisica, Università di Milano-Bicocca, P.zza della Scienza 3, 20126 Milano, Italy
[20] Institute of Physics, Jan Kochanowski University, ul. Swietokrzyska 15, 25-406 Kielce, Poland
[21] INFN, Sezione di Bologna, viale Berti Pichat 6/2, 40127 Bologna, Italy
[22] Institute d'Astrophysique de Paris, UMR7095 CNRS, Université Pierre et Marie Curie, 98bis Boulevard Arago, 75014 Paris, France
[23] Universitätssternwarte München, Ludwig-Maximillians Universität, Scheinerstr. 1, 81679 München, Germany
[24] Max-Planck-Institut für Extraterrestrische Physik, 84571 Garching b. München, Germany
[25] Institute of Cosmology and Gravitation, Dennis Sciama Building, University of Portsmouth, Burnaby Road, Portsmouth, PO1 3FX, UK
[26] INAF – Istituto di Astrofisica Spaziale e Fisica Cosmica Bologna, via Gobetti 101, 40129 Bologna, Italy
[27] INAF – Istituto di Radioastronomia, via Gobetti 101, 40129 Bologna, Italy
[28] Università degli Studi di Milano, via G. Celoria 16, 20130 Milano, Italy