Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **453,** 122–127 (2015) doi:10.1093/mnras/stv1617

A support vector machine for spectral classification of emission-line galaxies from the Sloan Digital Sky Survey

Fei Shi,★ Yu-Yan Liu,★ Guang-Lan Sun,★ Pei-Yu Li, Yu-Ming Lei and Jian Wang

North China Institute of Aerospace Engineering, Langfang, Hebei 065000, China

Accepted 2015 July 15. Received 2015 July 13; in original form 2015 February 21

ABSTRACT

The emission-lines of galaxies originate from massive young stars or supermassive black-holes. As a result, spectral classification of emission-line galaxies into star-forming galaxies, active galactic nucleus (AGN) hosts, or compositions of both relates closely to formation and evolution of galaxy. To find efficient and automatic spectral classification method, especially in large surveys and huge data bases, a support vector machine (SVM) supervised learning algorithm is applied to a sample of emission-line galaxies from the Sloan Digital Sky Survey (SDSS) data release 9 (DR9) provided by the Max Planck Institute and the Johns Hopkins University (MPA/JHU). A two-step approach is adopted. (i) The SVM must be trained with a subset of objects that are known to be AGN hosts, composites or star-forming galaxies, treating the strong emission-line flux measurements as input feature vectors in an n-dimensional space, where n is the number of strong emission-line flux ratios. (ii) After training on a sample of emission-line galaxies, the remaining galaxies are automatically classified. In the classification process, we use a 10-fold cross-validation technique. We show that the classification diagrams based on the [N II]/Hα versus other emission-line ratio, such as [O III]/Hβ, [Ne III]/[O II], ([O III]λ4959+[O III]λ5007)/[O III]λ4363, [O II]/Hβ, [Ar III]/[O III], [S II]/Hα, and [O I]/Hα, plus colour, allows us to separate unambiguously AGN hosts, composites or star-forming galaxies. Among them, the diagram of [N II]/Hα versus [O III]/Hβ achieved an accuracy of 99 per cent to separate the three classes of objects. The other diagrams above give an accuracy of ∼91 per cent.

Key words: methods: data analysis – galaxies: abundances – galaxies: starburst – galaxies: star formation.

1 INTRODUCTION

The emission line of galaxies originates from massive young stars or from a supermassive black hole. Accordingly, there are several existing types of emission-line galaxies: the two main classes are star-forming galaxies (SFGs) and active galactic nuclei (AGNs). Emission lines are observed in SFG because gas is ionized by massive and young stars. In contrast, AGN host galaxies contain a supermassive black hole, and their emission lines come from gas ionized by the light emitted from their accretion disc, according to the standard unified model (e.g. Antonucci 1993). A third class of emission-line galaxies show the characteristics on both kind of galaxies. As a result, spectral classification of emission-line galaxies into SFGs, AGN hosts, or a composition of both relates closely to formation and evolution of galaxy.

Furthermore, the study of dependence between the emission-line galaxy properties and the physical parameters, such as galaxy mass or environment can greatly benefit from the efficient classification of emission-line galaxies as AGN, composite, or SFG. The classification of different types of emission-line galaxies is one of the basic and crucial tasks to perform before moving on to any scientific analysis. To classify emission-line galaxies, the standard method used for the classification of AGNs and SFG in the local Universe is usually achieved by the use of the Baldwin, Phillips, Terlevich (BPT) diagnostic diagram (Baldwin, Phillips & Terlevich 1981). The BPT diagnostic diagram has been later revised and developed in many works. Among them, the most widely used are those by Kewley et al. (2001a,b, 2006) and Kauffmann et al. (2003). Veilleux & Osterbrock (1987) developed other diagnostic diagrams that involve [O III]/Hβ versus [S II]/Hα, and [S II]/Hα versus [O I]/Hα. Lamareille et al. (2004) and Lamareille (2010) established a classification using [O II]/Hβ versus [O II]/Hβ. Yan et al. (2011) derived a similar new diagnostic based on U-B rest-frame colours versus [O III]/Hβ.

★ E-mail: fshi@bao.ac.cn (FS); 275098078@qq.com (LY-Y); 1216606205@qq.com (G-LS)

Besides the widely used emission-line ratio above, there are other strong emission-line ratios in the optical band, such as for instance, [Ne III]/[O II], ([O III]λ4959+[O III]λ5007)/[O III]λ4363, [Ar III]/[O III]. These strong emission-line ratios have not been used to classify emission-line galaxies till now. Our aim is to study whether all these emission-line ratios in the optical band can be used to classify emission-line galaxies.

While classification by BPT diagnostic diagram is still common, there have been demand to use automated machine-learning techniques because astronomical data sets have grown considerably in size in the last decade owing largely to the advent of mosaic CCDs that can be used on large telescopes in order to image large areas of the sky down to very faint magnitudes. The Sloan Digital Sky Survey (SDSS; York et al. 2000) has led to the construction of a data set of over five million astronomical spectra since its first light in 1998. Future generations of wide-field imaging surveys such as the Dark Energy Survey,[1] PanStarrs[2] and Large Synoptic Survey Telescope (LSST)[3] will reach new limits in terms of the size of astronomical data sets. The automated classification algorithms will prove invaluable for the analysis of such data sets, but these algorithms are not yet to be applied on spectral classification of emission-line galaxies.

In this work, we employed an automatic support vector machine (SVM) classification of AGN, composite or SFG hosts, in the ninth Sloan Digital Sky Survey data release (SDSS DR9), by combining all strong emission-line flux ratio measurements including [Ne III]/[O II], [N II]/Hα, [O III]λλ4959,5007/[O III]λ4363, [S II]/Hα, [O II]λλ3726,3729/Hβ, [O III]λ5007/Hβ, Hα/Hβ, [Ar III]/[O III], and [O I]/Hα, provided by MPA/JHU. An SVM approach has already been successfully applied in astronomy for mainly the following problems: sorting out metal-poor galaxies (Shi et al. 2014) and quasar (Gao, Zhang & Zhao 2008; Peng et al. 2012), classification of variable stars (Woźniak et al. 2001, 2004), photometric classification of galaxies/AGNs/stars (Małek et al. 2013), morphological classification of galaxies (Humphreys et al. 2001; Huertas-Company et al. 2008, 2009), classification of multiwavelength data (Zhang & Zhao 2003, 2004), estimation of photometric redshifts of galaxies (Wadadekar 2005), matching different object catalogues in astrophysics (Rohde et al. 2005, 2006), and specific AGN subclass (BL Lacertae and flat-spectrum radio quasars; Hassan et al. 2013).

This paper is organized as follows. In Section 2, we describe the data set used for training and testing our analysis. We present a detailed description of our methodology in Section 3. We test the performance of our approach in Section 4 by applying to the data and present our results. We compare BPT classification with SVM in Section 5. Finally, we give our conclusions in Section 5. Throughout the paper, we adopt cosmological parameters $\Omega_{\rm M} = 0.27$ and $\Omega_{\Lambda} = 0.73$.

2 DATA SAMPLE

To use SVM, we must first construct a sample of sources with good emission-line detections. All the objects in the sample must have evident and reliable flux ratio measurements, such as [Ne III]/[O II], [N II]/Hα, [O III]λλ4959,5007/[O III]λ4363, [S II]/Hα, [O II]/Hβ, [O III]/Hβ, Hα/Hβ, [Ar III]/[O III], and [O I]/Hα. At the same time, all the objects in the sample must have accurate classification of AGN, composite, or SFG. For this purpose, we use the catalogue of galaxies in SDSS DR9 provided by MPA/JHU,[4] which made use of the spectral diagnostic diagrams from Kauffmann et al. (2003) to classify galaxies as AGN, composite, or SFG. In total, 113 336 galaxies are adopted in our sample, including 84 860 SFGs (74.9 per cent), 16 289 composites (14.4 per cent) and 12 187 AGN (10.7 per cent). All the galaxies in the sample have reliable spectral observations with reasonable values of strong line ratio and established classification as AGN, composite, or SFG. The redshifts of the galaxies in the sample are in the range of $0.02 < z < 0.3$.

3 SVM APPROACH

SVM are a particular family of machine learning technologies, first introduced by Vapnik (1995) to classify in a multidimensional parameter space. The idea of SVM is to map input vectors non-linearly into a high-dimensional feature space and construct the optimal decision hyperplane for classification in this high-dimensional feature space. As a result, the key problem of the SVM is to calculate decision hyperplanes between sets of objects having different classes.

The SVM implementation used is LIBSVM[5] (Chang & Lin 2011), an integrated software for SVM classification, which is currently one of the most widely used SVM software. The package is at the web site.[6] A typical use of LIBSVM involves two steps: first, training a data set that have known classifications to obtain a model determining the decision hyperplanes and second, using the model to predict information of a testing data set.

The data set of training sample includes the input vector $\boldsymbol{x}$ and target $\boldsymbol{t}$. The input vector $\boldsymbol{x}$ consists of redshift, [N II]/Hα, [O III]/Hβ, [Ne III]/[O II], ([O III]λ4959+[O III]λ5007)/[O III]λ4363, [O II]/Hβ, [Ar III]/[O III], [S II]/Hα, Hα/Hβ, and [O I]/Hα, plus colour (11 input variables) from the data set in Section 2. Target $\boldsymbol{t}$ is defined as 1, 2 or 3 to represent AGN, composite, or SFG.

To get the best performance, we use a 10-fold cross-validation technique. We first randomly divided the full sample into 10 subsets of equal size and selected 9 subsets to train the classification model. The remaining subset is used as a completely independent test of generalization. This test was repeated 10 times, with a different subset replaced for each training run. After completing the 10-fold cross-validation process, the classification accuracy was averaged over the 10 runs.

The LIBSVM algorithm need a non-linear kernel function mapping from the input space $\boldsymbol{x}$ to the feature space, so as to search for a hyperplane that maximize the distance from the boundary to the closest points belonging to the separate classes of objects. We choose a Gaussian radial basis kernel (RBK) function, which is one of the most popular SVM kernel functions, to make the non-linear feature map and defined as

$$k(x_{\rm i}, x_{\rm j}) = \exp(-\gamma||x_{\rm i} - x_{\rm j}||^2), \quad (1)$$

where $||x_{\rm i} - x_{\rm j}||$ is the Euclidean distance between each input variables $x_{\rm i}$ and $x_{\rm j}$, and $\gamma > 0$ parameter determines the topology of the decision surface. A low value of γ sets a very rigid, and complicated decision boundary and a high value of γ can give a very smooth decision surface causing misclassifications. Besides γ parameter, LIBSVM algorithm need a cost factor parameter C, that sets the width of the margin between hyperplanes separating different classes of objects. A large C value sets a small margin of separation between

[1] https://www.darkenergysurvey.org
[2] http://pan-starrs.ifa.hawaii.edu/public/
[3] http://www.lsst.org
[4] http://www.sdss3.org/dr9/spectro/spectroaccess.php
[5] Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm
[6] http://www.csie.ntu.edu.tw/∼cjlin/libsvm

different classes of objects and can lead to overfitting. A small C will make the hyperplane between different classes of objects smoother and can lead to some misclassifications.

To build a classifier that will be able to separate different classes of objects with good accuracy, it is necessary to tune the γ and C parameters. In order to find the most proper parameter γ and C, we performed a grid search with values from $\gamma \in (-4:4)$ and $C \in (-2:4)$ for every fold in the 10-fold cross-validation process.

A schematic representation of the SVM algorithm classification process can be summarized in the following steps.

(1) Choosing the *i*th training sample and the *i*th testing sample by cross-validation technique.
(2) Optimization of γ and C parameters the *i*th training sample and the *i*th testing sample.
(3) Training *i*th training sample.
(4) Classifying the *i*th testing sample.
(5) Moving to the $(i+1)$th training sample and the $(i+1)$th testing sample and go back to the step (2).
(6) After completing the 10 looped process above, the classification accuracy was averaged over the 10 runs.

4 SPECTRAL CLASSIFICATION OF GALAXIES

For illustration, we considered four SVM configurations that differ in terms of the number of variables. The first one uses all variables: redshift, [Ne III]/[O II], [N II]/Hα, [O III]$\lambda\lambda$4959,5007/[O III]λ4363, [S II]/Hα, [O II]/Hβ, [O III]/Hβ, Hα/Hβ, [Ar III]/[O III], and [O I]/Hα, plus colour. In the second configuration, we study strong line pairs, such as [N II]/Hα versus other emission-line ratios, one by one to show which line ratio is the most effective in identifying AGN, composite, or SFG. In the third configuration, we study how the classification accuracy depends on the internal reddening correction. In the fourth configuration, we study which line ratio can be useful to classify the emission-line galaxies when adding it to line ratio pair.

The confusion matrices of the first configuration is plotted in Fig. 1. For the introduction to confusion matrices, see the Matlab Recognizing Patterns web site.[7] For the first configuration using all variables, we achieved a classification accuracy of ~98.9 per cent. It is therefore apparent that the 11-variable SVM should be used for the purpose of selecting AGN, composite, or SFG in any optical spectral catalogue.

For the second configuration, we identified AGN, composite, or SFG only using the essential information for each strong line pairs, including redshift and Hα/Hβ line ratio. We found that the diagram of [N II]/Hα versus [O III]/Hβ achieved an accuracy of 98.8 per cent for classification of AGN hosts, composite or SFG. It is as a matter of course that the established classification for the galaxies in the sample is based on the diagram of [N II]/Hα versus [O III]/Hβ (Kauffmann et al. 2003). The SFGs of Kauffmann et al. (2003) are classified from their fig. 1, lying below the most conservative AGN rejection criterion. They are expected to have very low (<1 per cent) contribution to Hα from AGN. The AGN population consists of galaxies above the upper line in their fig. 1. This line corresponds to the theoretical upper limit for pure starburst models so that a substantial AGN contribution to the emission-line fluxes is required

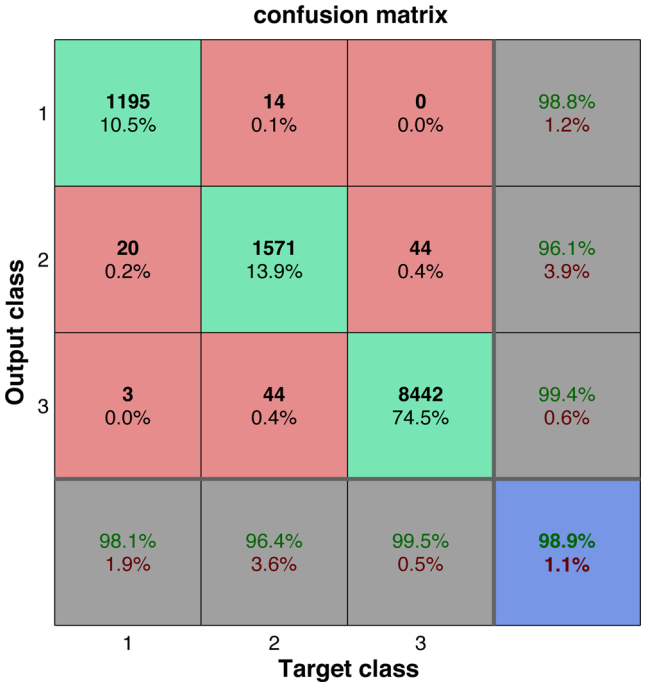

Figure 1. Confusion matrices using all 11 variables. The diagonal cells show the number of cases that were correctly classified, and the off-diagonal cells show the misclassified cases. The blue cell in the bottom right shows the total percent of correctly classified cases (in green) and the total percent of misclassified cases (in red). Target class 1, 2, 3 is the authentic classification of AGN, composite, or SFG and Output class 1, 2, 3 is the classification of AGN, composite, or SFG from SVM, respectively.

to move a galaxy above this line. The line is taken from equation 5 in Kewley et al. (2001a). The objects that are between the dotted and dashed curves in their fig. 1 are regarded as the composite galaxies.

The diagrams of [N II]/Hα versus other emission-line ratios, such as [Ne III]/[O II], ([O III]λ4959+[O III]λ5007)/[O III]λ4363, [O II]/Hβ, [Ar III]/[O III], [S II]/Hα, and [O I]/Hα, plus colour, give an accuracy of ~91 per cent (see Table 1). For the rest of strong line pairs, the classification accuracy worsens, therefore they cannot be used to identify AGN, composite, or SFG. For the diagrams between colour and strong emission lines except [N II]/Hα, the classification accuracy is also bad and cannot be used to identify AGN, composite, or SFG. In any case, it is an essential parameter for having redshift because it is vital to make the accurate redshift correction when deriving the flux of the strong emission line.

It is interesting to evaluate how well SVM will perform for the input vector [Ne III]/[O II], ([O III]λ4959+[O III]λ5007)/[O III]λ4363, [O II]/Hβ, [Ar III]/[O III], [S II]/Hα, and [O I]/Hα, plus colour (i.e. without the most useful inputs [N II]/Hα versus [O III]/Hβ) since the diagrams of [N II]/Hα versus [O III]/Hβ gives very similar results to that of using information of all emission-line ratios. The classification accuracy of that is 88.1 per cent which is less than the diagram using [N II]/Hα versus other emission-line ratio. So we conclude that [N II]/Hα is vital in the identification of AGN hosts, composites or SFGs.

For the third configuration, we remove Hα/Hβ line ratio from each strong line pairs to show the influence of extinction of interstellar dust on identifying AGN, composite or SFG(see Table 2). The extinction of interstellar dust in emission-line galaxies modifies the spectra of these objects. It is necessary to correct all observed line fluxes for this internal reddening in process of the study of the

[7] http://www.mathworks.com

Table 1. Classification accuracy for AGN, composite, or SFG as a function of the strong line pairs.

	[O III]/H β	[Ne III]/[O II]	*O*3ratio[a]	[O II]/H β	[Ar III]/[O III]	[S II]/H α	[O I]/H α	$g-r$
[N II]/H α	0.988	0.920	0.913	0.915	0.916	0.908	0.917	0.907
[O III]/H β		0.865	0.857	0.859	0.852	0.854	0.874	0.884
[Ne III]/[O II]			0.810	0.831	0.807	0.796	0.846	0.809
*O*3ratio[a]				0.775	0.781	0.774	0.835	0.818
[O II]/H β					0.784	0.781	0.827	0.799
[Ar III]/[O III]						0.778	0.835	0.799
[S II]/H α							0.831	0.786
[O I]/H α								0.837

[a]O3 ratio is ([O III]λ4959+[O III]λ5007)/[O III]λ4363.

Table 2. Classification accuracy for AGN, composite, or SFG depends on adding another line ratio to line ratio pairs.

	[O III]/H β	[Ne III]/[O II]	*O*3*ratio*	[O II]/H β	[Ar III]/[O III]	[S II]/H α	[O I]/H α	$g-r$
[N II]/H α[a]	0.988	0.919	0.917	0.910	0.915	0.907	0.913	0.909
[N II]/H α[b]	0.988	0.920	0.913	0.915	0.916	0.908	0.917	0.907

[a]Strong emission-line ratios removing Hα/Hβ to identify AGN, composite, or SFG.
[b]Strong emission-line ratios using Hα/Hβ to identify AGN, composite, or SFG.

dependence of emission-line galaxy properties on physical parameters such as metallicity, galaxy mass or environment. Hα/Hβ line ratio is probed because the relative strengths of low-order Balmer lines is the most widely used method to correct the emission-line spectra for the presence of dust (Shi et al. 2005; Shi, Kong & Cheng 2006). We find that the classification accuracy do not change when removing Hα/Hβ from them, which can be explained by the uncertainty of the internal reddening correction being comparable to the emission-line measurements and these flux ratios is not sensitive to the internal reddening correction because the wavelength separation between the two lines is small.

For the fourth configuration, we study which line ratio can be useful to classify the emission-line galaxies or be just noise by adding another line ratio to line ratio pairs (see Table 3). From Table 3, we can draw the conclusions as follows.

(1) In any case, adding [N II]/H α or [O III]/H β will greatly improve the classification accuracy.
(2) Adding [O I]/H α or $g-r$ colour will increase the classification accuracy of several percents in general for most line ratio pairs except [N II]/H α taking part in.
(3) Adding [Ne III]/[O II] will increase the classification accuracy of several percents in general for most line ratio pairs except [N II]/H α and [O III]/H β taking part in.
(4) Adding the rest of line ratios will not change the classification accuracy in general or even be just noise to decrease the classification accuracy.

5 COMPARE WITH BPT DIAGRAMS

We compare our result with the diagnostic from Kewley et al. (2006, hereafter K06) and Lamareille et al. (2010, hereafter L10). The result lists in Table 4.

The K06 diagrams use the following demarcation lines:

$$\log([\mathrm{O\,III}]/\mathrm{H}\beta) = 0.61/[\log([\mathrm{N\,II}]/\mathrm{H}\alpha) - 0.05] + 1.30, \quad (2)$$

where AGNs are above this curve which corresponds to the theoretical upper limit for pure starburst models, and

$$\log([\mathrm{O\,III}]/\mathrm{H}\beta) = 0.61/[\log([\mathrm{N\,II}]/\mathrm{H}\alpha) - 0.47] + 1.19, \quad (3)$$

with SFGs below this curve that lie below the most conservative AGN rejection criterion. Composites fall between these two curves.

When we use the K06 diagrams above to classify our sample, we find 10.7 per cent of the galaxies in the sample are classified as AGNs, 74.9 per cent of the galaxies in the sample are classified as SFGs, and the rests are composites, which is consistent with authentic classification(See Table 4). It is as a matter of course that authentic classification is derived by MPA/JHU, which use the similar classification method with K06 diagrams (Brinchmann et al. 2004).

The L10 diagrams use the following demarcation lines:

$$\log([\mathrm{O\,III}]/\mathrm{H}\beta) = 0.11/[\log([\mathrm{O\,II}]/\mathrm{H}\beta) - 0.92] + 0.85, \quad (4)$$

where AGNs are above this curve and SFGs are below this curve. Composites can be located by the two following inequalities:

$$\log([\mathrm{O\,III}]/\mathrm{H}\beta) \leq -(x-1)^2 - 0.1x + 0.25, \quad (5)$$

$$\log([\mathrm{O\,III}]/\mathrm{H}\beta) \geq (x-0.2)^2 - 0.60, \quad (6)$$

with $x = \log([\mathrm{O\,II}]/\mathrm{H}\beta)$.

When we use the L10 diagrams to classify our sample, we find 6626 (5.8 per cent of the total size) galaxies are classified as AGNs, 86 636 (76.4 per cent of the total size) of galaxies are classified as SFGs, and 20 061 (17.7 per cent of the total size) of galaxies are composites, which have a large deviation with SVM, especially for AGNs selection (See Table 4). The large deviation with SVM can be explained that many Seyfert 2 and LINERs will be omitted when using the equation 4 to select AGN and their composite galaxies are mixed with SFGs and AGNs regions (See fig. 4 of Lamareille et al. 2010).

6 CONCLUSIONS

We have presented a promising SVM approach to classify emission-line galaxies into AGN hosts, composites, or SFGs from spectral catalogues. The input variables are spectral measurements, i.e. redshift and the most observably strong emission-line ratios. Such multivariate analysis should be used in the classification of emission-line galaxies because some BPT diagrams using only two emission-line

Table 3. Classification accuracy for AGN, composite, or SFG depends on internal reddening correction.

	[N II]/H α	[O III]/H β	[Ne III]/[O II]	*O*3ratio	[O II]/H β	[Ar III]/[O III]	[S II]/H α	[O I]/H α	$g-r$
[N II]/H α, [O III]/H β	–	–	0.988	0.987	0.988	0.987	0.990	0.987	0.988
[N II]/H α, [Ne III]/[O II]	–	0.988	–	0.916	0.923	0.918	0.908	0.916	0.908
[N II]/H α, *O*3ratio	–	0.987	0.916	–	0.916	0.911	0.914	0.914	0.905
[N II]/H α, [O II]/H β	–	0.988	0.923	0.916	–	0.911	0.906	0.905	0.901
[N II]/H α, [Ar III]/[O III]	–	0.987	0.918	0.911	0.911	–	0.908	0.916	0.904
[N II]/H α, [S II]/H α	–	0.990	0.908	0.914	0.906	0.908	–	0.903	0.890
[N II]/H α, [O I]/H α	–	0.987	0.916	0.914	0.905	0.916	0.903	–	0.905
[N II]/H α, $g-r$	–	0.988	0.908	0.905	0.901	0.904	0.890	0.905	–
[O III]/H β, [Ne III]/[O II]	0.988	–	–	0.840	0.860	0.839	0.837	0.872	0.873
[O III]/H β, *O*3ratio	0.987	–	0.840	–	0.847	0.841	0.816	0.867	0.879
[O III]/H β, [O II]/H β	0.988	–	0.860	0.847	–	0.845	0.833	0.881	0.881
[O III]/H β, [Ar III]/[O III]	0.987	–	0.839	0.841	0.845	–	0.813	0.868	0.874
[O III]/H β, [S II]/H α	0.990	–	0.837	0.816	0.833	0.813	–	0.867	0.867
[O III]/H β, [O I]/H α	0.987	–	0.872	0.867	0.881	0.868	0.867	–	0.883
[O III]/H β, $g-r$	0.988	–	0.873	0.879	0.881	0.874	0.867	0.883	–
[Ne III]/[O II], *O*3ratio	0.916	0.840	–	–	0.836	0.823	0.817	0.860	0.837
[Ne III]/[O II], [O II]/H β	0.923	0.860	–	0.836	–	0.840	0.839	0.847	0.851
[Ne III]/[O II], [Ar III]/[O III]	0.918	0.839	–	0.823	0.840	–	0.812	0.858	0.830
[Ne III]/[O II], [S II]/H α	0.908	0.837	–	0.817	0.839	0.812	–	0.836	0.820
[Ne III]/[O II], [O I]/H α	0.916	0.872	–	0.860	0.847	0.858	0.836	–	0.865
[Ne III]/[O II], $g-r$	0.908	0.873	–	0.837	0.851	0.830	0.820	0.865	–
*O*3ratio, [O II]/H β	0.916	0.847	0.836	–	–	0.749	0.767	0.854	0.834
*O*3ratio, [Ar III]/[O III]	0.911	0.841	0.823	–	0.749	–	0.788	0.842	0.832
*O*3ratio, [S II]/H α	0.914	0.816	0.817	–	0.767	0.788	–	0.845	0.822
*O*3ratio, [O I]/H α	0.914	0.867	0.860	–	0.854	0.842	0.845	–	0.861
*O*3ratio, $g-r$	0.905	0.879	0.837	–	0.834	0.832	0.822	0.861	–
[O II]/H β, [Ar III]/[O III]	0.911	0.845	0.840	0.749	–	–	0.784	0.846	0.836
[O II]/H β, [S II]/H α	0.906	0.833	0.839	0.767	–	0.784	–	0.843	0.829
[O II]/H β, [O I]/H α	0.905	0.881	0.847	0.854	–	0.846	0.843	–	0.855
[O II]/H β, $g-r$	0.901	0.881	0.851	0.834	–	0.836	0.829	0.855	–
[Ar III]/[O III], [S II]/H α	0.908	0.813	0.812	0.788	0.784	–	–	0.822	0.825
[Ar III]/[O III], [O I]/H α	0.916	0.868	0.858	0.842	0.846	–	0.822	–	0.860
[Ar III]/[O III], $g-r$	0.904	0.874	0.830	0.832	0.836	–	0.825	0.860	–
[S II]/H α, [O I]/H α	0.903	0.867	0.836	0.845	0.843	0.822	–	–	0.840
[S II]/H α, $g-r$	0.890	0.867	0.820	0.822	0.829	0.825	–	0.840	–
[O I]/H α, $g-r$	0.905	0.883	0.865	0.861	0.855	0.860	0.840	–	–

Table 4. The percent of classified as AGNs, Composites and SFGs in the sample derived by BPT diagrams and SVM.

	AGNs	Composites	SFGs
K06	0.107	0.144	0.749
L10	0.058	0.177	0.764
SVM	0.105	0.139	0.745
Authentic[a]	0.107	0.144	0.749

[a]The percent of AGNs, Composites and SFGs in the sample provided by MPA/JHU is regarded as the authentic per cent.

ratios, for example L10, tend to misclassify a considerable part of AGNs.

In the target classification, we achieved a classification accuracy of 98.8 per cent using the diagram of [N II]/Hα versus [O III]/Hβ which means that classification to be either an AGN hosts galaxy, a composite, or an SFG has a 98.8 per cent chance to be correct. The classification accuracy will decrease to ∼91 per cent if using [N II]/Hα versus other emission-line ratio, such as [Ne III]/[O II], ([O III]λ4959+[O III]λ5007)/[O III]λ4363, [O II]/Hβ, [Ar III]/[O III], [S II]/Hα, and [O I]/Hα, plus colour. All the diagrams are reliable to a certain degree if including [N II]/Hα. It shows a serious potential to search for new AGN or SFG candidate with [N II]/Hα versus other emission-line ratio when [O III]/Hβ is unavailable and K06 classification is infeasible.

This new statistical method developed in the context of the SDSS project can be extended easily to any other analysis requiring AGN or SFG selection when the physical property of the target can be quantitative. The code in the paper is available on the web.[8]

Finally, we note that, aside from its relative simplicity and robustness, the SVM classification method that we presented here can be extended and improved in a number of ways, such as X-ray or radio selected AGN, photometric selection of AGN or SFG, or making more detailed AGN classification into Seyfert galaxies, Quasars, LINERs, and so on. One has to be cautioned that both the classification accuracy and run time may change dramatically in these processes.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (NSFC, Nos.11203001, 11202003, 11225315,

[8] http://fshi5388.blog.163.com

1320101002, 11433005, and 11421303), the Strategic Priority Research Program 'The Emergence of Cosmological Structures' of the Chinese Academy of Sciences (No. XDB09000000), the Specialized Research Fund for the Doctoral Program of Higher Education (SRFDP, No. 20123402110037), the natural science foundation of Hebei Province (No. A2014409002) and the Chinese National 973 Fundamental Science Programs (973 program) (2015CB857004).

Funding for the SDSS has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Aeronautics and Space Administration, the National Science Foundation, the US Department of Energy, the Japanese Monbukagakusho, and the Max Planck Society.

REFERENCES

Antonucci R., 1993, ARA&A, 31, 473
Baldwin J. A., Phillips M. M., Terlevich R., 1981, PASP, 93, 5
Brinchmann J., Charlot S., White S. D. M., Tremonti C., Kauffmann G., Heckman T., Brinkmann J., 2004, MNRAS, 351, 1151
Chang C.-C., Lin C.-J., 2011, ACM Trans. Intelligent Systems and Technology, 2, 27
Gao D., Zhang Y.-X., Zhao Y.-H., 2008, MNRAS, 386, 1417
Hassan T., Mirabal N., Contreras J. L., Oya I., 2013, MNRAS, 428, 220
Huertas-Company M., Rouan D., Tasca L., Soucail G., Le Fèvre O., 2008, A&A, 478, 971
Huertas-Company M. et al., 2009, A&A, 497, 743
Humphreys R. M., Karypis G., Hasan M., Kriessler J., Odewahn S. C., 2001, BASS, 33, 1322
Kauffmann G. et al., 2003, MNRAS, 346, 1055
Kewley L. J., Heisler C. A., Dopita M. A., Lumsden S., 2001a, ApJS, 132, 37
Kewley L. J., Dopita M. A., Sutherland R. S., Heisler C. A., Trevena J., 2001b, ApJ, 556, 121
Kewley L. J., Groves B., Kauffmann G., Heckman T., 2006, MNRAS, 372, 961 (K06)
Lamareille F., 2010, A&A, 509, A53 (L10)
Lamareille F., Mouhcine M., Contini T., Lewis I., Maddox S., 2004, MNRAS, 350, 396
Małek K. et al., 2013, A&A, 557, A16
Peng N., Zhang Y., Zhao Y., Wu X.-b., 2012, MNRAS, 425, 2599
Rohde D. J., Drinkwater M. J., Gallagher M. R., Downs T., Doyle M. T., 2005, MNRAS, 360, 69
Rohde D. J., Gallagher M. R., Drinkwater M. J., Pimbblet K. A., 2006, MNRAS, 369, 2
Shi F., Kong X., Li C., Cheng F. Z., 2005, A&A, 437, 849
Shi F., Kong X., Cheng F. Z., 2006, A&A, 453, 487
Shi F., Liu Y.-Y., Kong X., Chen Y., Li Z.-H., Zhi S.-T., 2014, MNRAS, 444, L49
Vapnik V. N., 1995, The Nature of Statistical Learning Theory. Springer, Berlin
Veilleux S., Osterbrock D. E., 1987, ApJS, 63, 295
Wadadekar Y., 2005, PASP, 117, 79
Woźniak P. R. et al., 2001, BAAS, 33, 1495
Woźniak P. R., Williams S. J., Vestrand W. T., Gupta V., 2004, AJ, 128, 2965
Yan R. et al., 2011, ApJ, 728, 38
York D. G. et al., 2000, AJ, 120, 1579
Zhang Y., Zhao Y., 2003, PASP, 115, 1006
Zhang Y., Zhao Y., 2004, A&A, 422, 1113

This paper has been typeset from a TEX/LATEX file prepared by the author.