Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **503,** 5263–5273 (2021) doi:10.1093/mnras/stab744

Classification of 4XMM-DR9 sources by machine learning

Yanxia Zhang ,[1]⋆ Yongheng Zhao[1] and Xue-Bing Wu[2,3]

[1] *CAS Key Laboratory of Optical Astronomy, National Astronomical Observatories, Beijing 100101, China*
[2] *Department of Astronomy, School of Physics, Peking University, Beijing 100871, China*
[3] *Kavli Institute for Astronomy and Astrophysics, Peking University, Beijing 100871, China*

Accepted 2021 March 9. Received 2021 February 22; in original form 2020 May 27

ABSTRACT

The ESA's *X-ray Multi-mirror Mission* (*XMM–Newton*) created a new high-quality version of the *XMM–Newton* serendipitous source catalogue, 4XMM-DR9, which provides a wealth of information for observed sources. The 4XMM-DR9 catalogue is correlated with the Sloan Digital Sky Survey (SDSS) DR12 photometric data base and the AllWISE data base; we then get X-ray sources with information from the X-ray, optical, and/or infrared bands and obtain the *XMM–WISE*, *XMM*–SDSS, and *XMM–WISE*–SDSS samples. Based on the large spectroscopic surveys of SDSS and the Large Sky Area Multi-object Fiber Spectroscopic Telescope (LAMOST), we cross-match the *XMM–WISE*–SDSS sample with sources of known spectral classes, and obtain known samples of stars, galaxies, and quasars. The distribution of stars, galaxies, and quasars as well as all spectral classes of stars in 2D parameter space is presented. Various machine-learning methods are applied to different samples from different bands. The better classified results are retained. For the sample from the X-ray band, a rotation-forest classifier performs the best. For the sample from the X-ray and infrared bands, a random-forest algorithm outperforms all other methods. For the samples from the X-ray, optical, and/or infrared bands, the LogitBoost classifier shows its superiority. Thus, all X-ray sources in the 4XMM-DR9 catalogue with different input patterns are classified by their respective models that are created by these best methods. Their membership of and membership probabilities for individual X-ray sources are assigned. The classified result will be of great value for the further research of X-ray sources in greater detail.

Key words: methods: data analysis – methods: statistical – astronomical data bases: miscellaneous – catalogues – stars: general – galaxies: general.

1 INTRODUCTION

Since all X-rays are prevented from entering by the Earth's atmosphere, only a space-based telescope can observe and probe celestial X-ray sources. Both NASA's *Chandra X-ray Observatory* and the ESA's *X-ray Multi-mirror Mission* (*XMM–Newton*) are space missions in the X-ray band and are leading X-ray astronomy into a new era (Brandt & Hasinger 2005). Significant discoveries have been made with these missions (Santos-Lleo et al. 2009). These missions may provide answers to other profound cosmic questions such as the enigmatic black holes, the formation and evolution of galaxies, dark matter, dark energy, the origins of the Universe, and so on. They are valuable tools to probe X-ray emission from various astrophysical systems. Despite the implementation of these missions, more and more X-ray sources have still not been identified. Identification of deep X-ray survey sources is a challenging issue for several reasons (Brandt & Hasinger 2005). Large sky-survey projects (e.g. the Sloan Digital Sky Survey, SDSS; the *Wide-field Infrared Survey Explorer*, *WISE*; the Large Sky Area Multi-object Fiber Spectroscopic Telescope, LAMOST) provide multiwavelength information and spectroscopic classes of X-ray sources. Pineau et al. (2011) cross-correlated the 2XMMi catalogue with SDSS DR7 and studied the high-energy properties of various classes of X-ray sources. Machine learning can gain knowledge from the known examples and create a classifier to predict unknown sources. Therefore machine learning makes it possible to classify X-ray sources depending on their multiwavelength and spectroscopic information for known samples. Some work has been done in this direction. For example, Broos et al. (2011) applied a naive Bayes classifier to classify X-ray sources from the *Chandra* Carina Complex Project. Zhang et al. (2013) ran a random-forest algorithm on the cross-matched sample between 2XMMi-DR3 and SDSS-DR8. Farrell, Murphy & Lo (2015) classified the variable 3XMM sources with the random-forest algorithm. Arnason, Barmby & Vulic (2020) identified new X-ray binary candidates in M31 also using the random-forest algorithm.

In this paper, we download the 4XMM-DR9 catalogue, and obtain the spectroscopic classes of these X-ray sources from SDSS and LAMOST, X-ray information from *XMM–Newton*, optical information from SDSS, and infrared information from AllWISE. We create classifiers to classify the X-ray sources with known spectroscopic classes based on only X-ray information; combined X-ray and optical/infrared information; or combined X-ray, optical, and infrared information. Section 2 describes the data used and the distribution of various objects in 2D space. Section 3 presents the classification methodologies. Section 4 compares the performance of better classifiers for different samples. Section 5 discusses the results

⋆ E-mail: zyx@bao.ac.cn

Published by Oxford University Press on behalf of Royal Astronomical Society

of the classifiers and applies the created classifiers to the unknown sources. Section 6 provides our conclusions for this work.

2 THE DATA

The European Space Agency's (ESA) *X-ray Multi-mirror Mission* (*XMM–Newton*) was launched on 1999 December 10, performing in the X-ray, ultraviolet, and optical bands. *XMM–Newton* is ESA's second cornerstone of the Horizon 2000 Science Programme. It carries three high-throughput X-ray telescopes with an unprecedented effective area and an optical monitor, the first flown on an X-ray observatory. This mission has released a new high-quality version of the *XMM–Newton* serendipitous source catalogue 4XMM-DR9. This catalogue includes 810 795 detections of 550 124 unique sources drawn from 11 204 *XMM–Newton* EPIC observations, covering 1152 degrees^2 of the sky in the energy band from 0.2–12 keV (Webb et al. 2020). For the total photon energy band from 0.2–12 keV, the median flux of the catalogue detections is $\sim 2.3 \times 10^{-14}$ erg cm^{-2} s^{-1}; it is $\sim 5.3 \times 10^{-15}$ erg cm^{-2} s^{-1} in the soft energy band (0.2–2 keV), and $\sim 1.2 \times 10^{-14}$ erg cm^{-2} s^{-1} in the hard band (2–12 keV). About 23 per cent of the sources have total fluxes below 1×10^{-14} erg cm^{-2} s^{-1}. The typical positional accuracy is about 2 arcsec. For the astrometric quality, the mean RA and Dec. offsets between the *XMM* sources and the SDSS optical quasars are -0.01 and 0.005 arcsec respectively with corresponding standard deviations of 0.70 and 0.64 arcsec (see fig. 10 in Webb et al. 2020).

The *Wide-field Infrared Survey Explorer* (*WISE*; Wright et al. 2010) is an entire mid-infrared sky survey with simultaneous photometry in four filters at 3.4, 4.6, 12, and 22 μm (*W*1, *W*2, *W*3, and *W*4). It obtained over a million images and observed hundreds of millions of celestial objects. The *WISE* survey provides mid-infrared information about the Solar system, the Milky Way, and the Universe. On the basis of the *WISE* work, the AllWISE programme has created new products with better photometric sensitivity and accuracy as well as better astrometric precision. The limiting magnitudes of *W*1 and *W*2 are brighter than 19.8 and 19.0 (Vega: 17.1, 15.7) for the AllWISE source catalogue. Sources brighter than 8, 7 in the *W*1 and *W*2 bands are affected by saturation. Considering the accuracy of *W*1, *W*2, *W*3, and *W*4, we only adopt *W*1 and *W*2, converting *W*1 and *W*2 in Vega magnitudes to AB magnitudes by $W1_{\rm AB} = W1 + 2.699$ and $W2_{\rm AB} = W2 + 3.339$. The average point spread function (PSF) with full widths at half-maximum (FWHMs) in *W*1, *W*2, *W*3, and *W*4 is 6.1, 6.4, 6.5, and 12 arcsec, respectively. For high signal-to-noise ratios (S/N) (>20) sources, the *WISE* positions are better than 0.15 arcsec for 1σ and one axis.

The Sloan Digital Sky Survey (SDSS; York et al. 2000) has been one of the most successful photometric and spectroscopic sky surveys ever made, providing deep multicolour images of a third of the sky and spectra for more than 3000 000 celestial objects. Data Release 12 (DR12) is the final data release of SDSS-III, containing all SDSS observations up to 2014 July (Eisenstein et al. 2011). It includes the complete data set of the BOSS and APOGEE surveys, and now also includes stellar radial velocity measurements from MARVELS. Data Release 16 (DR16) is the fourth SDSS data release (SDSS-IV; Blanton et al. 2017). SDSS mapped the sky in the five optical band passes (*ugriz*) with central wavelengths of 3551, 4686, 6165, 7481, and 8931 Å. Pixel size is 0.396 arcsec and the astrometry accuracy is less than 0.1 arcsec rms absolute per coordinate. The limiting magnitudes of *ugriz* are 21.6, 22.2, 22.2, 21.3, and 20.7 at 95 per cent completeness, respectively. For *u* and *z*, they are converted to AB magnitudes by $u_{\rm AB} = u - 0.04$ mag and $z_{\rm AB} = z + 0.02$ mag. DR16 contains SDSS observations up to 2018 August, including 880 652 stars, 2616 381 galaxies, and 749 775 quasars when *zWarning* = 0 in the DR16 SpecObj data base.

The Large Sky Area Multi-object Fiber Spectroscopic Telescope (LAMOST; Cui et al. 2012; Luo et al. 2015) may take 4000 spectra in a single exposure to a limiting magnitude as faint as $r = 19$ mag at the resolution $R = 1800$. It has finished the first five-year survey plan. The LAMOST survey contains the LAMOST ExtraGAlactic Survey (LEGAS) and the LAMOST Experiment for Galactic Understanding and Exploration (LEGUE) survey of Milky Way stellar structure. The data products of the fifth data release (DR5) include 8183 160 stars (7540 605 stars with S/N in the *g* or *i* bands greater than 10), 152 863 galaxies, 52 453 quasars, and 637 889 unknown objects.

The SDSS Data Release 14 Quasar catalogue (DR14Q; Pâris et al. 2018) contains 526 356 spectroscopically identified quasars. DR14Q consists of spectroscopically identified quasars from SDSS-I, II, III, and the latest SDSS-IV eBOSS survey.

In order to obtain multiwavelength properties of X-ray sources, we cross-match the 4XMM-DR9 catalogue with the SDSS and AllWISE data bases. According to the work of Covey et al. (2008), we estimate spurious SDSS and AllWISE matches by applying a 30 arcsec offset to the X-ray source declinations and searching the 4XMM-DR9 catalogue for sources with SDSS counterparts within 8 arcsec and AllWISE counterparts within 10 arcsec for each X-ray source centroid. Fig. 1 shows a normalized cumulative histogram of separation between the 4XMM-DR9 and SDSS sources as well as the 4XMM-DR9 and AllWISE sources; the solid histogram represents the cumulative distribution of separation between the X-ray and optical counterparts for real *XMM*–SDSS sources within 8 arcsec (left-hand panel of Fig. 1) and that of separation between the X-ray and infrared counterparts for real *XMM–WISE* sources within 10 arcsec (right-hand panel of Fig. 1); the dashed histogram indicates the upper limit to the fractional contamination of the *XMM*–SDSS sample by chance superpositions of independent X-ray and optical sources (left-hand panel of Fig. 1) and the *XMM–WISE* sample by chance superpositions of independent X-ray and infrared sources (right-hand panel of Fig. 1). In general, high completeness and low contamination cannot be achieved at the same time; higher completeness is needed at the expense of contamination; otherwise, if we pursue low contamination, we must sacrifice completeness. For *XMM* matching SDSS at 3, 4, 5, and 6 arcsec, the completeness versus contamination is respectively 71.68 per cent versus 9.75 per cent, 79.49 per cent versus 16.53 per cent, 85.55 per cent versus 24.50 per cent, and 90.78 per cent versus 33.36 per cent; for *XMM* matching AllWISE at 3, 6, 7, and 8 arcsec, the completeness versus contamination is respectively 56.90 per cent versus 0.89 per cent, 77.66 per cent versus 1.75 per cent, 83.39 per cent versus 2.01 per cent, and 89.14 per cent versus 2.08 per cent. From Fig. 1, the fraction of X-ray sources matching SDSS occupies over 90 per cent at 6 arcsec and that matching AllWISE is about 90 per cent at 8 arcsec. So the cross-match radius between the SDSS and 4XMM-DR9 sources is set as 6 arcsec while that between AllWISE and 4XMM-DR9 sources is adopted as 8 arcsec. We apply the software TOPCAT (Taylor 2005) to perform the cross-matching. Finally we obtain the *XMM–WISE* sample and the *XMM*–SDSS sample; the *XMM–WISE*–SDSS sample is then derived according to the same ID (srcid) in the *XMM–WISE* and *XMM*–SDSS samples. All photometries throughout this paper are extinction-corrected according to the work of Schindler et al. (2017) and AB magnitudes are adopted.

In order to construct the known spectral samples, the samples have been identified spectroscopically by SDSS DR16 and LAMOST DR5. The known samples are cross-matched with the *XMM–WISE*–

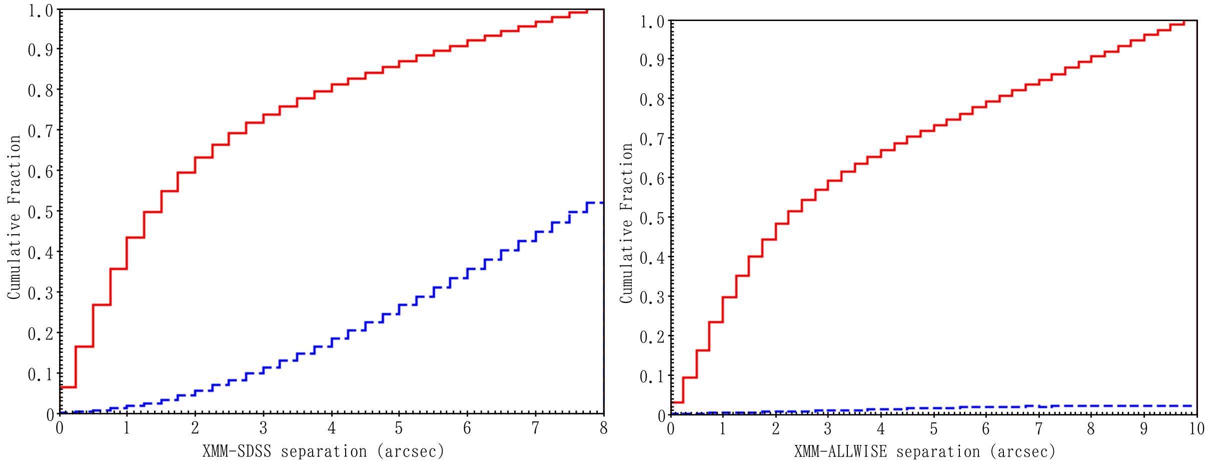

Figure 1. Histogram of source separation between *XMM* and SDSS as well as between *XMM* and AllWISE. Red solid histogram: cumulative distribution of separation between X-ray and optical counterparts (left) and between X-ray and infrared counterparts (right) for the real X-ray sources; blue dashed line: distribution of separations returned by matching the faked X-ray sources with coordinates shifted by 30 arcsec to the SDSS (left) and AllWISE (right) data bases.

SDSS sample in a 6 arcsec radius. Keeping the data quality, *zWarning* = 0 is set in the DR16 SpecObj data base when downloading data; sc_poserr≤5 and sc_sum_flag<3 are set in the 4XMM-DR9 data base; records with default values of *ugriz*, *W*1, and *W*2 are removed; records with $W1 < 8$ and $W2 < 7$ are deleted; and stars in the LAMOST DR5 data base are adopted with S/N in the *g* or *i* bands greater than 10. When the objects are both identified by SDSS and LAMOST, only the spectral class of objects in SDSS are retained. If the objects with known spectral class in the *XMM–WISE–*SDSS sample have counterparts in DR14Q, the objects are labelled as QSO. Finally, the known samples include 3558 stars, 7203 galaxies, and 21 040 quasars with information from the X-ray, optical, and infrared bands. The spectra were identified as stars, galaxies, and QSO by the SDSS and LAMOST automated classification pipelines using template fitting. Detailed information on known samples is given in Table 1. For the class assigned as galaxies, the subclass Non is from the LAMOST data base and the subclass's default value is from the SDSS data base. The LAMOST pipeline does not provide subclasses for galaxies, and all the subclasses for galaxies in the LAMOST data base are labelled as Non. The websites relating to the above data sets are shown in Table 2. As for the definitions and abbreviations in Table 1, AGN is short for active galactic nucleus, AGN BL for broad-line AGN, SB for starburst galaxy, SB BL for broad-line SB, SF for star-forming galaxy, SF BL for broad-line SF, BL for BL Lacertae objects, CV for cataclysmic variable star, EM for emission line star, WD for white dwarf, DB for double or binary star, sdM1 for subdwarf M1 star, Carbon for carbon star, and O, B, A, F, G, K, M for stars with spectral types of O, B, A, F, G, K, M, respectively. All these subclasses are assigned by the SDSS and LAMOST automated classification pipelines depending on the spectroscopic characteristics. Bolton et al. (2012) showed that the galaxy spectra from SDSS by the line-fitting code were grouped into AGN, SF, and SB; if the spectra meet log10([O III]/H β) >1.2 log10([N II]/H α) + 0.22, the galaxy spectra were identified as AGN, otherwise, for the equivalent width (EW) of H α, SF if EW(H α) <50 Å, and SB if EW(H α) >50 Å; galaxies and quasars may be classified as broad-line (BL) when their line widths are larger than 200 km s^{-1}; and stellar spectra were classified as spectral types from O to M based on the ELODIE stellar library. The broad-line classification given by the SDSS pipeline does not necessarily indicate that an AGN has emission lines broad enough to classify it as a broad-line (as opposed to a narrow-line) AGN because the emission line widths are typically more than 2000 km s^{-1} for broad-line AGNs (Hao et al. 2005). BL Lacertae objects are a subclass of AGNs that have fast and large amplitude variability over the whole

Table 1. The numbers of each class and subclass for the known samples.

Class	Subclass	No.
Galaxy	AGN	611
	AGN BL	107
	SB	387
	SB BL	8
	SF	1008
	SF BL	46
	BL	281
	Non	219
		4536
Star	O	1
	B	5
	A	79
	F	708
	G	869
	K	777
	M	1062
	CV	39
	DB	5
	EM	1
	WD	10
	sdM1	1
	Carbon	1
QSO		21 040

Table 2. The websites for related catalogues.

4XMM-DR9 catalogue
https://www.cosmos.esa.int/web/xmm-newton/xsa
Spectrally identified stars, galaxies and quasars from SDSS
http://skyserver.sdss.org/dr16/en/tools/search/sql.aspx
Spectrally identified stars, galaxies, and quasars from LAMOST
http://dr5.lamost.org/v3/catalogue
SDSS DR14 Quasar catalogue (DR14Q)
https://www.sdss.org/dr14/algorithms/qso_catalog

Table 3. The parameters, definition, catalogues, and wavebands.

Parameter	Definition	Catalogue	Waveband
srcid	Source ID	*XMM*	X-ray band
sc_ra	Right ascension in decimal degrees	*XMM*	X-ray band
sc_dec	Declination in decimal degrees	*XMM*	X-ray band
$hr1$	Hardness ratio 1 Definition: $hr1 = (B - A)/(B + A)$, where A = count rate in energy band 0.2–0.5 keV B = count rate in energy band 0.5–1 keV	*XMM*	X-ray band
$hr2$	Hardness ratio 2 Definition: $hr2 = (C - B)/(C + B)$, where B = count rate in energy band 0.5–1 keV C = count rate in energy band 1–2 keV	*XMM*	X-ray band
$hr3$	Hardness ratio 3 Definition: $hr3 = (D - C)/(D + C)$, where C = count rate in energy band 1–2 keV D = count rate in energy band 2–4.5 keV	*XMM*	X-ray band
$hr4$	hardness ratio 4 Definition: $hr4 = (E - D)/(E + D)$, where D = count rate in energy band 2–4.5 keV E = count rate in energy band 4.5–12 keV	*XMM*	X-ray band
extent	Source extent	*XMM*	X-ray band
$\log(f_{\rm x})$	X-ray flux	*XMM*	X-ray band
$\log(f_{\rm x}/f_{\rm r})$	X-ray-to-optical-flux ratio	SDSS, *XMM*	Optical and X-ray bands
u	u magnitude	SDSS	Optical band
g	g magnitude	SDSS	Optical band
r	r magnitude	SDSS	Optical band
i	i magnitude	SDSS	Optical band
z	z magnitude	SDSS	Optical band
$W1$	$W1$ magnitude	AllWISE	Infrared band
$W2$	$W2$ magnitude	AllWISE	Infrared band

spectra, high and variable polarization, and continuous spectra with no or weak absorption and emission features. Starburst galaxies are characterized by higher rates of star formation than normal galaxies. They are either young or rejuvenated galaxies that typically contain very luminous X-ray sources. Since the separation of subclasses of galaxies depends on spectral line information, it is difficult to discriminate them without spectra.

We select the features [$\log(f_x)$, $hr1$, $hr2$, $hr3$, $hr4$, *extent*, r, $W1$, $u - g$, $g - r$, $r - i$, $i - z$, $z - W1$, $W1 - W2$, $\log(f_x/f_r)$] of the known star, galaxy, and quasar samples used for this study. The selected features are described in Table 3. The 2D plots between two attributes from these features are given in Figs 2 and 3. Fig. 2 shows the differences between stars, galaxies, and quasars, while Fig. 3 indicates the differences between different spectral classes of stars. The two figures tell us that it is difficult to discriminate stars, galaxies, and quasars, and different spectral classes of stars depending only on two attributes. These attributes all contribute more or less to the classification. As shown in Fig. 2, most quasars obviously have larger $\log(f_x/f_r)$, r, $W1$, and $W1 - W2$ values than stars. It is easy to classify stars and quasars from galaxies with the attribute *extent* in the X-ray band. Nevertheless some AGNs do not appear as X-ray extended if the emission is nuclear-dominated; thus they are misclassified as stars or quasars only depending on *extent*. We check stars and quasars with large *extent* in SIMBAD and NED within a 3 arcsec radius, and find that some of them are galaxies in a group of galaxies, galaxy cluster, or other kinds of objects. Most galaxies indeed have a relatively larger *extent* in the X-ray band. Most galaxies overlap with most quasars in their X-ray and infrared information while most galaxies overlap with most stars in their X-ray, optical, and infrared information. In order to effectively separate stars, galaxies, and quasars, it is necessary to apply all available information. As indicated in Fig. 3, CV stars have more strong X-ray emission than other stars, and most CV stars and M stars have more strong infrared emission than the remainder of the stars. They can be separated easily from the star sample in some 2D spaces. Apparently they have obvious differences from the remainder of the star sample as they are mixed together and are thus difficult to discriminate.

3 THE METHOD

WEKA (the Waikato Environment for Knowledge Analysis; Witten & Frank 2005) is a piece of open-source software that is effectively used for various machine-learning tasks. It is implemented through a graphical user interface, standard terminal applications, or through a JAVA API. It is widely used for teaching, research, and industrial applications, and contains a plethora of built-in tools for standard machine-learning tasks. These tasks include data pre-processing, classification, regression, clustering, association rules, attribute selection, and visualization realized by different algorithms. This software makes it easy to work with large amounts of data and to run and compare various machine-learning algorithms. It has been successfully applied in astronomy (Zhao & Zhang 2008; Zhang, Zhao & Gao 2008; Zheng & Zhang 2008).

We try various classification algorithms provided by WEKA on our samples and only keep the better classification results. When running the software, we all adopt the default setting by 10-fold validation while training a model. 10-fold validation refers to a data set that is randomly divided into 10 parts, nine parts of which are used for training with one part remaining for testing; this procedure is repeated 10 times.

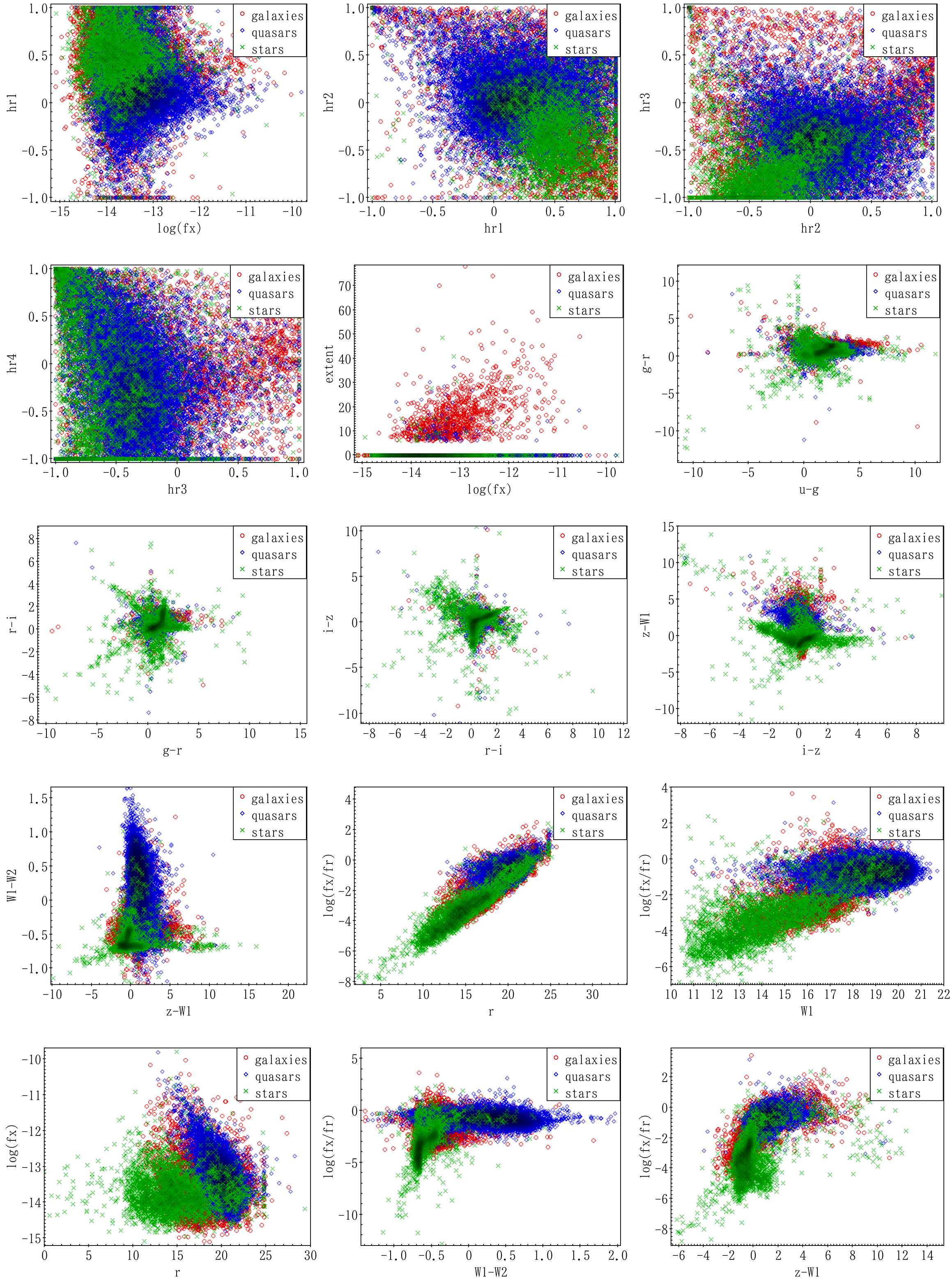

Figure 2. The distribution of stars, galaxies, and quasars in 2D space; red open circles represent galaxies, blue open diamonds represent quasars, and green crosses represent stars.

The metrics commonly used to evaluate the performance of a classifier include accuracy, precision, recall, and F-measure. For a data set, accuracy is the ratio of the total number of correct predictions to the total number of predictions, precision (also called efficiency) is the fraction of true positive predictions among all true positive examples, recall (also called completeness) is the fraction of true positive predictions among all predicted positive examples, and F-measure is the weighted average of precision and

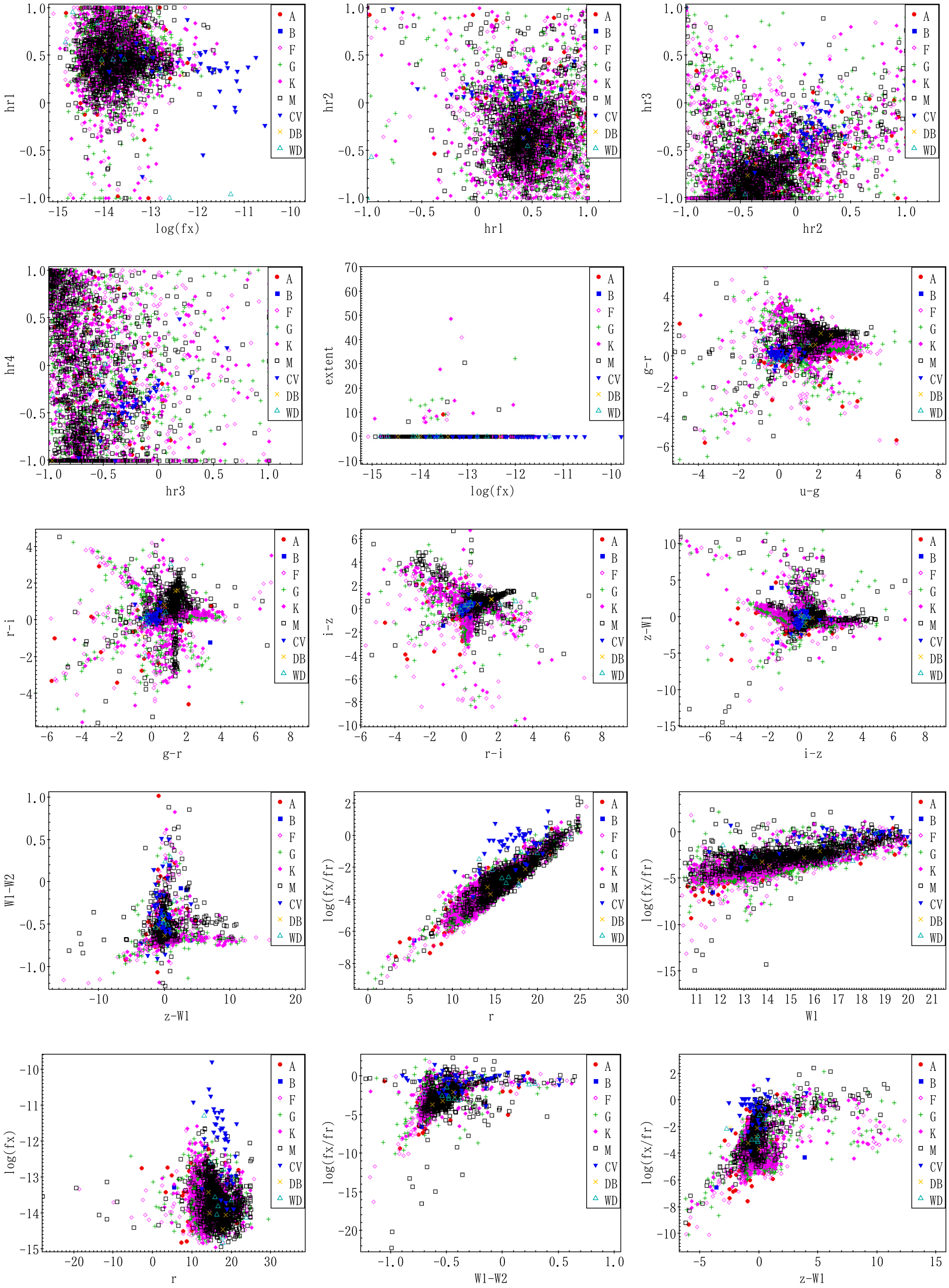

Figure 3. The distribution of different spectral classes of stars in 2D space; red filled circles for A stars, blue filled squares for B stars, purple open diamonds for F stars, green pluses for G stars, purple filled diamonds for K stars, black open squares for M stars, blue filled down triangles for CV stars, yellow crosses for double stars (DB), and light-green open up triangles for white dwarf stars (WD).

recall:

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{1}$$

Here TP is the true positive sample, TN is the true negative sample, FP is the false positive sample, FN is the false negative sample:

$$\text{precison} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \; \text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{F-measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \tag{3}$$

3.1 Random forest

Random forest is a supervised learning algorithm that builds a randomized decision tree in each iteration of the bagging algorithm and gives impressive results with very large ensembles (Breiman 2001). The bagging algorithm is applied to improve accuracy by reducing the variance to make the model more general and avoiding overfitting. For bagging, multiple subsets is taken as the training set. For each subset, a model created by the same algorithm is used to predict the output for the same test set. Averaging predictions is considered as the final prediction output. To further understand how the bagging algorithm works, we assume that there are *N* models and a data set. This data set is split into training and test sets. Taking a sample of records from the training set, we train the first model with it. Then, taking another sample from the training set, we train the second model with it. A similar process will be repeated for *N* number of models. Based on all predictions of *N* models on the same test set, we adopt a model-averaging technique like weighted average, variance, or max voting to obtain the final prediction. Ensembles are a divide-and-conquer approach used to improve performance. For ensemble methods, 'weak learners' are grouped to form a 'strong learner'. Each classifier individually is a 'weak learner' (base learner) while all the classifiers taken together are a 'strong learner'. In a decision tree, the input data are separated into smaller and smaller sets from the tree root to its leaves. A random forest creates many decision trees. When classifying a new object, each decision tree provides a classification. The final class of this object depends on the most votes among all the trees in the forest. This simplified random forest is shown in Fig. 4. The advantage of using random forest is that it is able to deal with unbalanced and missing data and runs relatively fast.

3.2 Rotation forest

Rotation forest is a powerful tree-based ensemble method based on feature extraction and is designed to work with a smaller number of ensembles; it focuses on building accurate and diverse classifiers (Rodriguez, Kuncheva & Alonso 2006). Feature extraction by principal component analysis (PCA) is performed on *K* subsets randomly split from the feature set in turn; here *K* is a rotation-forest parameter. All principal components are kept for each subset. The original data are handled by the principal component transformation and then used for training each base classifier. Its diversity is realized by the feature extraction carried out on each base classifier and its accuracy is ensured by the keeping of all principal components and the use of all of the data as a training sample for each base classifier. Decision trees are usually selected because they are easily influenced by rotation of the feature axes. The difference between random forest and rotation forest is that rotation forest performs PCA on the feature subset to rebuild full feature space and achieves similar or better performance with fewer trees than random forest does. For detailed information on rotation forest refer to Rodriguez et al. (2006).

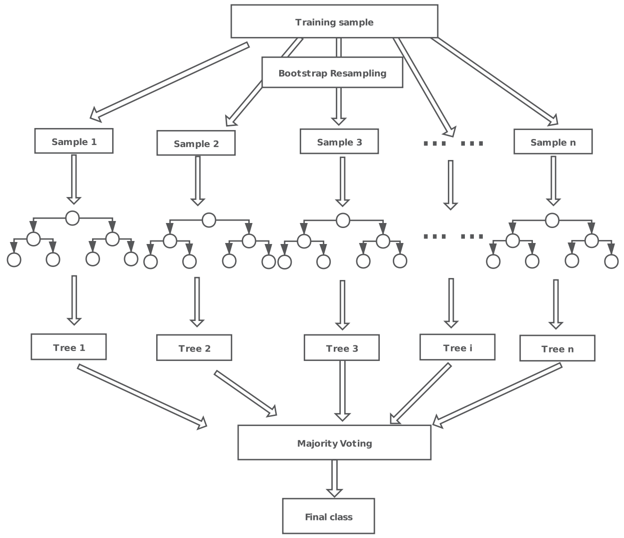

Figure 4. The simplified random forest.

3.3 LogitBoost

LogitBoost is a boosting classification algorithm, based on the logistic regression method by minimizing the logistic loss (Friedman, Hastie & Tibshirani 2000). Because noise and outliers exist in data and an exponential loss function is used in LogitBoost, issues like overfitting will reduce the model accuracy. However classification errors are changed linearly instead of exponentially; thus this may improve the model accuracy and noise immunity. Here the LogitBoost classification algorithm is trained using random forests as weak learners.

4 PERFORMANCE OF THE ALGORITHMS

We classify the X-ray sources into some subclasses of galaxies, stars, and quasars, based on the input pattern of $\log(f_x)$, $hr1$, $hr2$, $hr3$, $hr4$, *extent*, r, $W1$, $u - g$, $g - r$, $r - i$, $i - z$, $z - W1$, $W1 - W2$, $\log(f_x/f_r)$. Since the LAMOST data base does not give the subclassification of galaxies, we do not consider galaxies from LAMOST when performing multiclassification. The subclasses of AGN and AGN BL are labelled as AGN; SB and SB BL as SB; SF and SF BL as SF; and the default value as galaxies. LogitBoost is applied to the known sample without the galaxies from LAMOST by 10-fold validation. The classified result is described in Table 4. As shown in Table 4, the total accuracy adds up to 90.04 per cent; the metrics of stars and quasars are above 92.7 per cent while those of galaxies are unsatisfactory. The subclasses of galaxies are easily confused. The subclass of default value for galaxies assigned as galaxy belongs to normal galaxies, while the subclasses of AGN, SF, SB, and BL belong to active galaxies. All metrics of normal galaxies are larger than 77.0 per cent while those of active galaxies range from 7.8 per cent to 76.6 per cent. Active galaxies are likely to be classified as normal galaxies or quasars. Obviously it is very difficult to discriminate active galaxies from the whole sample. Therefore we use the known samples from LAMOST and SDSS, and only classify the sample into galaxies, stars, and quasars in the following work.

Table 4. The performance of LogitBoost for multiclassification.

Known↓Classified→	AGN	BL	SB	SF	Galaxy	QSO	Star	Precision	Recall	F-measure
AGN	149	8	7	98	269	177	10	50.0%	20.8%	29.3%
BL	12	22	1	9	193	34	10	44.9%	7.8%	13.3%
SB	4	0	141	68	23	141	18	76.6%	35.7%	48.7%
SF	53	3	28	472	251	205	42	58.9%	44.8%	50.9%
Galaxy	45	14	4	100	3698	605	70	77.0%	81.5%	79.2%
QSO	35	0	0	50	268	20 657	30	94.0%	98.2%	96.1%
Star	0	2	3	4	103	148	3298	94.8%	92.7%	93.7%
Total accuracy					90.04%					

For the sample from the X-ray band only, the classification performance of random forest and rotation forest is shown in Table 5. The input pattern for this sample is log(f_x), *hr*1, *hr*2, *hr*3, *hr*4, *extent*. As shown in Table 5, for galaxies only, recall and F-measure decrease, but for stars and quasars, all metrics increase, comparing the performance of rotation forest with random forest. Rotation forest outperforms random forest in terms of accuracy (77.80 per cent versus 77.46 per cent). With information from the X-ray band only, the classification metrics of quasars are satisfying while those of galaxies and stars are not good when considering precision, recall, and F-measure.

For the sample from the X-ray and optical bands, the classification performance of random forest and LogitBoost is indicated in Table 6. The input pattern is log(f_x), *hr*1, *hr*2, *hr*3, *hr*4, *extent*, r, $u - g$, $g - r$, $r - i$, $i - z$, log(f_x/f_r). As indicated in Table 6, all metrics for LogitBoost are better than those for random forest, and all of them are higher than 84.8 per cent. Only touching on quasars and stars, the metrics are above 87.5 per cent. LogitBoost is superior to random forest for this case, as its accuracy amounts to 92.82 per cent.

For the sample from the X-ray and infrared bands, the classification performance of random forest and LogitBoost is described in Table 7. The input pattern is log(f_x), *hr*1, *hr*2, *hr*3, *hr*4, *extent*, $W1$, $W1 - W2$. As depicted in Table 7, the performance of random forest is a little better than LogitBoost in terms of total accuracy. All metrics for random forest are near to those of LogitBoost. The accuracy of galaxies is still worse than that of quasars and stars. Nevertheless, all metrics are better than 76.1 per cent. The total accuracy of random forest is 89.42 per cent.

For the sample from the X-ray, optical, and infrared bands, the classification performance of random forest and LogitBoost is listed in Table 8. The input pattern is log(f_x), *hr*1, *hr*2, *hr*3, *hr*4, *extent*, r, $W1$, $u - g$, $g - r$, $r - i$, $i - z$, $z - W1$, $W1 - W2$, log(f_x/f_r). As shown in Table 8, even for galaxies, the metrics are greater than 87.1 per cent; for stars, the metrics are above 90.7 per cent; for quasars, the metrics are higher than 95.4 per cent. All metrics except precision for LogitBoost are greater than those of random forest. Compared to random forest, LogitBoost has a slight advantage and its total accuracy adds up to 94.26 per cent.

In order to check how the observational errors influence the performance of a classifier, we take the *XMM*–SDSS sample as an example. Setting $\sigma_u < 0.3$, $\sigma_g < 0.3$, $\sigma_r < 0.3$, $\sigma_i < 0.3$, and $\sigma_z < 0.3$, the known sample size changes from 31 800 to 26 428; the performance of random forest and LogitBoost is shown in Table 9. Comparing the result in Table 9 with that in Table 6, the performances of random forest and LogitBoost both improve with higher-quality data (94.73 per cent versus 92.57 per cent for random forest, 94.93 per cent versus 92.82 per cent for LogitBoost) in terms of accuracy. Although higher-quality data lead to higher performance of a classifier, the number of sources with X-ray emission is small in nature, so we do not set a magnitude error limitation on the samples in our work.

5 DISCUSSION AND APPLICATION

Comparing Tables 5–8, the worst result belongs to the sample from the X-ray band only, as expected. Adding the information from the optical and/or infrared bands, the classification accuracy increases for any classifier; nevertheless the accuracy with the X-ray and optical bands is better than that with the X-ray and infrared bands. The best performance is obtained with all information from the X-ray, optical, and infrared bands. There is no algorithm that shows the best performance for every data set. For the sample from the X-ray band, the rotation-forest classifier is the best; for the sample from the X-ray and infrared bands, random forest is superior to all other algorithms; and for another two samples, LogitBoost shows its superiority.

In reality, some X-ray sources have information from the X-ray, optical, and infrared bands, some have information from the X-ray and infrared bands, some have information from the X-ray and optical bands, and some even have X-ray information only. Based on the known samples with spectral classes, we need to construct four classifiers for the four situations to predict the unknown X-ray sources. For the sources with X-ray information only, a rotation-forest classifier is built with the known samples with spectral classes to predict their classes and probability. For the sources with X-ray and infrared bands, a random-forest classifier is created with the known samples with spectral classes to predict their classes and probability. For the sources from the X-ray and optical bands or from the X-ray, optical, and infrared bands, LogitBoost classifiers are constructed with the corresponding known samples with spectral classes to predict their classes and probability, respectively. For the 4XMM-DR9 sources, all predicted results are shown in Table 10. Table 10 provides classification information for the 4XMM-DR9 sources. The information gained will be of great value for further research into the characteristics and physics of X-ray sources.

6 CONCLUSIONS

Based on the distribution of stars, galaxies, and quasars in 2D space, it is difficult to discriminate them and their subclasses clearly. Similarly, given the distribution of all spectral classes of stars in 2D space, it is also not so easy to separate them, but CV stars and M stars stand out clearly in some 2D spaces. Of the entire X-ray sample, quasars occupy the majority while stars and galaxies only cover a minority. With X-ray information and spectral classes of known X-ray sources, we create a rotation-forest classifier to

Table 5. The performance of random forest and rotation forest with $\log(f_x)$, $hr1$, $hr2$, $hr3$, $hr4$, *extent*.

Method		Random forest			Rotation forest	
Class	Precision	Recall	F-measure	Precision	Recall	F-measure
QSO	82.1%	93.4%	87.4%	81.4%	94.9%	87.6%
Galaxy	63.0%	43.4%	51.4%	66.0%	40.4%	50.1%
Star	64.0%	52.4%	57.6%	65.1%	52.7%	58.3%
Total accuracy		77.46%			77.80%	

Table 6. The performance of random forest and LogitBoost with $\log(f_x)$, $hr1$, $hr2$, $hr3$, $hr4$, *extent*, r, $u - g$, $g - r$, $r - i$, $i - z$, $\log(f_x/f_r)$.

Method		Random forest			LogitBoost	
Class	Precision	Recall	F-measure	Precision	Recall	F-measure
QSO	94.4%	96.1%	95.2%	94.5%	96.2%	95.4%
Galaxy	85.6%	84.8%	85.2%	86.1%	85.3%	85.7%
Star	95.9%	87.5%	91.5%	96.2%	88.1%	92.0%
Total accuracy		92.57%			92.82%	

Table 7. The performance of random forest and LogitBoost with $\log(f_x)$, $hr1$, $hr2$, $hr3$, $hr4$, *extent*, $W1$, $W1 - W2$.

Method		Random forest			LogitBoost	
Class	Precision	Recall	F-measure	Precision	Recall	F-measure
QSO	93.4%	95.9%	94.6%	93.4%	96.8%	95.9%
Galaxy	79.1%	76.1%	77.5%	78.9%	76.1%	76.1%
Star	85.1%	78.2%	81.5%	85.1%	82.3%	77.9%
Total accuracy		89.42%			89.38%	

Table 8. The performance of random forest and LogitBoost with $\log(f_x)$, $hr1$, $hr2$, $hr3$, $hr4$, *extent*, r, $W1$, $u - g$, $g - r$, $r - i$, $i - z$, $z - W1$, $W1 - W2$, $\log(f_x/f_r)$.

Method		Random forest			LogitBoost	
Class	Precision	Recall	F-measure	Precision	Recall	F-measure
QSO	95.4%	97.0%	96.2%	95.6%	97.1%	96.3%
Galaxy	88.4%	87.1%	87.7%	89.1%	87.5%	88.3%
Star	96.9%	90.7%	93.7%	96.8%	91.2%	93.9%
Total accuracy		94.03%			94.26%	

Table 9. The performance of random forest and LogitBoost with $\log(f_x)$, $hr1$, $hr2$, $hr3$, $hr4$, *extent*, r, $u - g$, $g - r$, $r - i$, $i - z$, $\log(f_x/f_r)$ when $\sigma_u < 0.3$, $\sigma_g < 0.3$, $\sigma_r < 0.3$, $\sigma_i < 0.3$, and $\sigma_z < 0.3$.

Method		Random forest			LogitBoost	
Class	Precision	Recall	F-measure	Precision	Recall	F-measure
QSO	95.7%	98.1%	96.9%	95.9%	98.1%	97.0%
Galaxy	88.8%	83.7%	86.2%	89.3%	84.1%	86.6%
Star	96.4%	89.8%	93.0%	96.5%	90.9%	93.6%
Total accuracy		94.73%			94.93%	

assign classification results and their probabilities for all 4XMM-DR9 sources. Based on information from the X-ray and infrared bands as well as spectral classes of known X-ray sources, a random-forest classifier is used to discriminate X-ray sources. By means of properties from the X-ray, optical, and/or infrared bands and spectral classes of known X-ray sources, we build LogitBoost classifiers to predict X-ray sources. The predicted results from different methods with different input properties are listed in full in a table, which may be used for further study of the X-ray properties of various kinds of objects in detail.

Table 10. Classification of 4XMM-DR9 sources.

srcid	sc_ra	sc_dec	Class_x	P_x	Class_xo	P_{xo}	Class_xi	P_{xi}	Class_xio	P_{xio}
200001101010001	64.9255899382624	55.9993455276706	galaxy	0.718	star	1.0				
200001101010002	64.9714038006107	55.8049026564271	galaxy	0.427	QSO	0.996	galaxy	0.894	QSO	0.998
200001101010003	65.0767247976311	55.9307646652894	galaxy	0.456	QSO	0.963	galaxy	1.0	star	0.722
200001101010004	65.1112285547752	55.9955363739078	galaxy	0.746	galaxy	1.0	galaxy	0.993	galaxy	1.0
200001101010005	64.996228987918	56.2248168838265	star	0.653	star	1.0	star	1.0	star	1.0
200001101010006	64.9348515102436	55.9291776566485	galaxy	0.506						
200001101010007	64.8232313435949	55.9849189955416	galaxy	0.485	QSO	1.0	galaxy	1.0	QSO	0.999
200001101010008	65.0734121719342	55.9823011754657	QSO	0.491	star	1.0	star	0.939	star	1.0
200001101010009	65.0167233356568	55.9421102139164	QSO	0.537	QSO	0.997				
200001101010010	64.9101008917805	56.0710218248335	QSO	0.502						
200001101010011	64.9050705152553	56.0644078750126	QSO	0.539	galaxy	0.955	galaxy	0.999	galaxy	0.943
200001101010012	65.2336693528087	55.8993466831422	galaxy	0.479			star	0.986		
200001101010013	64.8914247247132	55.9585111145714	QSO	0.523						
200001101010014	64.6507013015166	56.0418886508129	QSO	0.517	star	0.999	galaxy	0.986	star	1.0
200001101010015	64.7925428495702	55.896051166999	star	0.572	star	1.0	star	1.0	star	1.0
200001101010016	65.1527613793266	55.9300031359814	QSO	0.489			galaxy	0.999		
200001101010017	65.0424438892887	56.1513807784794	QSO	0.488	QSO	1.0				
200001101010018	64.725085117142	55.891398599223	galaxy	0.579	star	1.0	galaxy	1.0	star	0.996
200001101010019	65.1553866221519	55.8977634034868	galaxy	0.452	galaxy	1.0	star	0.654	star	0.702
200001101010020	64.9468670845107	55.9626521430764	galaxy	0.657	galaxy	0.981	galaxy	0.998	galaxy	0.98

Notes. Class_x means classification and P_x shows their classification probabilities from the X-ray band; Class_xo means classification and P_{xo} shows their classification probabilities from the X-ray and optical bands; Class_xi means classification and P_{xi} shows their classification probabilities from the X-ray and infrared bands; Class_xio means classification and P_{xio} shows their classification probabilities from the X-ray, infrared, and optical bands.
This whole table is available at http://paperdata.china-vo.org/zyx/table10.csv. Part of it is shown here to demonstrate its form and content.

ACKNOWLEDGEMENTS

We are very grateful to the referees for their constructive suggestions. This paper is funded by the National Natural Science Foundation of China under grants No. 11873066 and No. U1731109. This research has made use of data obtained from the 4XMM *XMM–Newton* serendipitous source catalogue compiled by the 10 institutes of the *XMM–Newton* Survey Science Centre selected by ESA. This publication makes use of data products from the *Wide-field Infrared Survey Explorer*, which is a joint project of the University of California, Los Angeles, and the Jet Propulsion Laboratory/California Institute of Technology, funded by the National Aeronautics and Space Administration. The Guoshoujing Telescope (the Large Sky Area Multi-object Fiber Spectroscopic Telescope, LAMOST) is a National Major Scientific Project built by the Chinese Academy of Sciences. Funding for the project has been provided by the National Development and Reform Commission. LAMOST is operated and managed by the National Astronomical Observatories, Chinese Academy of Sciences.

We acknowledgment SDSS data bases. Funding for the Sloan Digital Sky Survey IV has been provided by the Alfred P. Sloan Foundation, the US Department of Energy Office of Science, and the Participating Institutions. SDSS-IV acknowledges support and resources from the Center for High-Performance Computing at the University of Utah. The SDSS web site is www.sdss.org. SDSS-IV is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration including the Brazilian Participation Group, the Carnegie Institution for Science, Carnegie Mellon University, the Chilean Participation Group, the French Participation Group, the Harvard-Smithsonian Center for Astrophysics, Instituto de Astrofísica de Canarias, the Johns Hopkins University, Kavli Institute for the Physics and Mathematics of the Universe (IPMU)/University of Tokyo, Lawrence Berkeley National Laboratory, Leibniz Institut für Astrophysik Potsdam (AIP), Max-Planck-Institut für Astronomie (MPIA Heidelberg), Max-Planck-Institut für Astrophysik (MPA Garching), Max-Planck-Institut für Extraterrestrische Physik (MPE), National Astronomical Observatories of China, New Mexico State University, New York University, University of Notre Dame, Observatário Nacional/MCTI, the Ohio State University, Pennsylvania State University, Shanghai Astronomical Observatory, the United Kingdom Participation Group, Universidad Nacional Autónoma de México, University of Arizona, University of Colorado Boulder, University of Oxford, University of Portsmouth, University of Utah, University of Virginia, University of Washington, University of Wisconsin, Vanderbilt University, and Yale University.

DATA AVAILABILITY

The predicted 4XMM-DR9 catalogue is available in a repository and can be accessed using a unique identifier; part of it is shown in Table 10. It is available on paperdata at http://paperdata.china-vo.org, and can be accessed at http://paperdata.china-vo.org/zyx/table10.csv.

REFERENCES

Arnason R. M., Barmby P., Vulic N., 2020, MNRAS, 492, 5075
Blanton M. R. et al., 2017, AJ, 154, 28
Bolton A. S. et al., 2012, AJ, 144, 144
Brandt W. N., Hasinger G., 2005, ARA&A, 43, 727
Breiman L., 2001, Machine Learning, 45, 5
Broos P. S., Getman K. V., Povich M. S., Townsley L. K., 2011, ApJS, 194, 4
Covey K. R. et al., 2008, ApJS, 178, 339
Cui X.-Q. et al., 2012, RAA, 12, 1197
Eisenstein D. J. et al., 2011, AJ, 142, 72
Farrell S. A., Murphy T., Lo K. K., 2015, ApJ, 813, 28
Friedman J., Hastie T., Tibshirani R., 2000, Ann. Statistics, 28, 337

Hao L. et al., 2005, AJ, 129, 1783

Luo A. L. et al., 2015, RAA, 15, 1095

Pâris I. et al., 2018, A&A, 613, A51

Pineau F.-X., Motch C., Carrera F., Della Ceca R., Derrière S., Michel L., Schwope A., Watson M. G., 2011, A&A, 527, A126

Rodriguez J. J., Kuncheva L. I., Alonso C. J., 2006, IEEE Trans. Pattern Analysis and Machine Intelligence, 28, 1619

Santos-Lleo M., Schartel N., Tananbaum H., Tucker W., Weisskopf M. C., 2009, Nature, 462, 997

Schindler J. T., Fan X., McGreer I. D., Yang Q., Wu J., Jiang L., Green R., 2017, ApJ, 851, 13

Taylor M. B., 2005, in Shopbell P., Britton M., Ebert R., eds, ASP Conf. Ser. Vol. 347, Astronomical Data Analysis Software and Systems XIV. Astron. Soc. Pac., San Francisco, p. 29

Webb N. A. et al., 2020, A&A, 641, A136

Witten I. H., Frank E., 2005, Data Mining: Practical Machine Learning Tools, Techniques with Java Implementations. Morgan Kaufmann, San Francisco

Wright E. L. et al., 2010, AJ, 140, 1868

York D. G. et al., 2000, AJ, 120, 1579

Zhang Y., Zhao Y., Gao D., 2008, Advances Space Res., 41, 1949

Zhang Y., Zhou X., Zhao Y., Wu X., 2013, AJ, 145, 42

Zhao Y., Zhang Y., 2008, Advances Space Res., 41, 1955

Zheng H., Zhang Y., 2008, Advances Space Res., 41, 1960

SUPPORTING INFORMATION

Supplementary data are available at *MNRAS* online.

Table 10. Classification of 4XMM-DR9 sources.

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

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