Statistical Challenges in 21st Century Cosmology
Proceedings IAU Symposium No. 306, 2014
A. F. Heavens, J.-L. Starck & A. Krone-Martins, eds.

doi:10.1017/S1743921315000022

Statistical analysis of cross-correlation sample of 3XMM-DR4 with SDSS-DR10 and UKIDSS-DR9

Yan-Xia Zhang[1], Yong-Heng Zhao[1], Xue-Bing Wu[2] and Hai-Jun Tian[1,3]

[1]Key Laboratory of Optical Astronomy, National Astronomical Observatories, Chinese Academy of Sciences, 20A Datun Road, Chaoyang District, 100012, Beijing, P.R.China
[2]Department of Astronomy, Peking University 100871, Beijing, P.R.China
[3]College of Science, China Three Gorges University, 443002 Yichang, P.R.China
email:zyx@bao.ac.cn

Abstract. We match the XMM-Newton 3XMMi-DR4 catalog with the Sloan Digital Sky Survey (SDSS) Data Release 10 and the United Kingdom Infrared Deep Sky Survey (UKIDSS) Data Release 9. Based on this X-ray/optical/infrared catalog, we probe the distribution of various types of X-ray emitters in the multidimensional parameter space. It is found that quasars, galaxies and stars have some kind distribution rule, especially for stars. The result shows that only using the X-ray/optical features, stars are difficult to discriminate from galaxies and quasars, the added information from infrared band is very helpful to improve the classification result of any classifier. Comparing the classification accuracy of random forests with that of rotation forests, rotation forests show better performance.

Keywords. catalogs - surveys - X-rays: diffuse background - X-rays: galaxies - X-rays: stars

1. Introduction

Various large sky surveys from different bands provide abundant resources of studying multiwavelength properties of objects. X-ray satellites are helpful to explore some types of objects, notably active galaxies (AGN), clusters of galaxies, interacting compact binaries and active stellar coronae. Zhang *et al.* (2013) cross-correlated the XMM-Newton 2XMMi-DR3 catalog with the Sloan Digital Sky Survey (SDSS) Data Release 8, investigated the distribution of various classes of X-ray emitters in the multidimensional photometric parameter space, then applied random forests on this sample and found that the X-ray emitting stars had poor classification accuracy compared to galaxies and quasars.

2. Data

3XMM-DR4 is generated from the European Space Agency's (ESA) XMM-Newton observatory. The catalogue contains 531,261 X-ray source detections above the processing likelihood threshold of 6. These X-ray source detections relate to 372,728 unique X-ray sources.

The Sloan Digital Sky Survey (SDSS) experiences over eight years of operations (SDSS-I, 2000-2005; SDSS-II, 2005-2008), now SDSS enters SDSS-III. Data Release 10 (DR10) is the first release of the spectra from the SDSS-III's Apache Point Observatory Galactic Evolution Experiment (APOGEE). DR10 also includes hundreds of thousands of new galaxy and quasar spectra from the Baryon Oscillation Spectroscopic Survey (BOSS), in addition to all imaging and spectra from prior SDSS data releases.

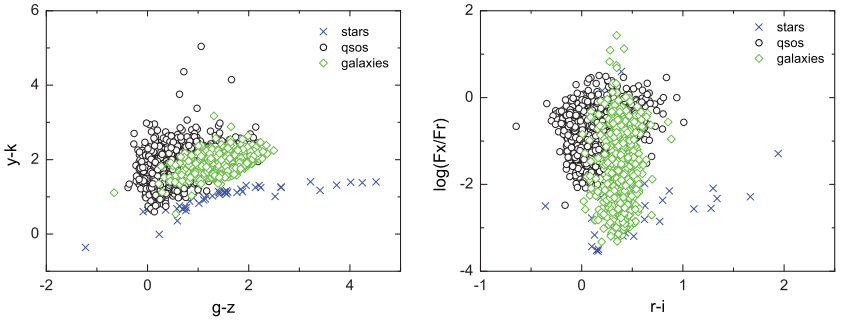

Figure 1. Left: Distribution of sources in $y-k$ vs. $g-z$ diagram. Right: Distribution of sources in the $\log(f_x/f_r)$ vs. $r-i$ diagram. Stars are represented as crosses, galaxies as open diamonds, quasars as circles.

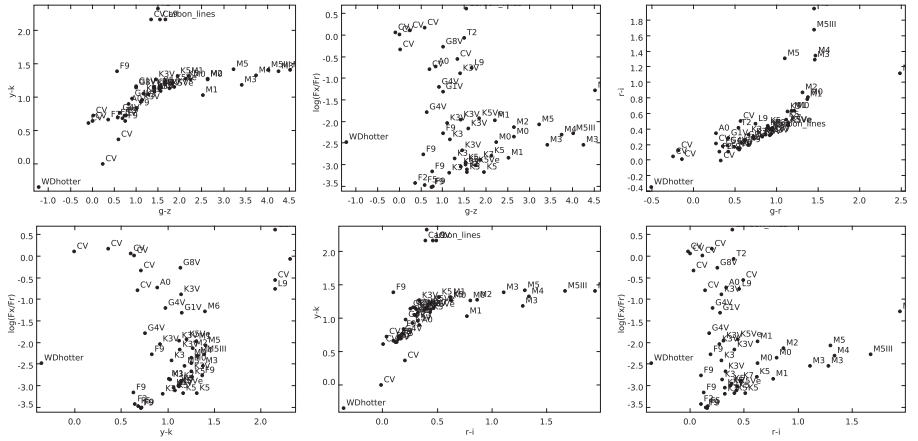

Figure 2. Distributions of stars in different parameter spaces.

The UKIRT InfraRed Deep Sky Survey (UKIDSS) consists of five surveys: Large Area Survey (LAS), Galactic Plane Survey (GPS), Galactic Clusters Survey, Deep Extragalactic Survey (DXS), Ultra Deep Survey (UDS). As for LAS, the number of objects observed is 212,234,393, and each band depth is Y=20.2, J=19.6, H=18.8, K=18.2, respectively. Here we use the public data release 9.

We cross-match 3XMM-DR4 with SDSS DR10 and UKIDSS DR9, and obtain the sample with information from X-ray, optical and infrared bands. In order to study statistical properties of this sample, Figures 1-2 show the distribution of quasars, galaxies and various stars in the multidimensional space.

3. Methods and Results

Random forests, developed by Leo Breiman and Adele Cutler, are one kind of decision tree methods by constructing a lot of decision trees and outputting the class that is the mode of the classes provided by individual trees. They have wide applications in astronomy for classification and regression.

Rotation forests (Rodriguez *et al.* 2006) are a method for creating classifier ensembles, based on feature extraction by Principal Component Analysis (PCA). Decision trees are adopted here. Rotation forests are firstly used in astronomy.

Using the open source software WEKA (Hall *et al.* 2009) and adopting the default setting, random forests and rotation forests are performed on the cross-matched sample.

https://doi.org/10.1017/S1743921315000022 Published online by Cambridge University Press

Table 1. Accuracy with different input patterns

Input pattern	Random Forest Accuracy	Rotation Forest Accuracy
$g-z, y-k$	87.00%	88.70%
$i, g-z, y-k$	88.34%	88.83%
$i, g-z, y-k, log(f_X/f_R)$	90.27%	91.03%
$hr1, hr2, hr3, hr4, i, g-z, y-k, log(f_X/f_R)$	91.12%	91.12%
$hr1, hr2, hr3, hr4, SC\_EXTENT, i, g-z, y-k, log(f_X/f_R)$	90.31%	91.21%
$u-g, g-r, r-i, i-z$	88.30%	89.60%
$i, u-g, g-r, r-i, i-z$	89.64%	91.30%
$r, u-g, g-r, r-i, i-z$	90.27%	91.21%
$y-j, j-h, h-k$	85.38%	87.94%
$y, y-j, j-h, h-k$	90.36%	91.79%
$u-g, g-r, r-i, i-z, y-j, j-h, h-k$	91.39%	92.42%
$i, u-g, g-r, r-i, i-z, y-j, j-h, h-k$	92.12%	**93.50%**
$i, u-g, g-r, r-i, i-z, y, y-j, j-h, h-k$	92.47%	93.05%
$u-g, g-r, r-i, i-z, g-z, y-j, j-h, h-k, y-k$	92.11%	93.00%
$i, u-g, g-r, r-i, i-z, g-z, y-j, j-h, h-k, y-k$	92.15%	93.27%
$i, u-g, g-r, r-i, i-z, log(f_X/f_R), y-j, j-h, h-k$	**92.65%**	93.41%
$SC\_EXTENT, i, u-g, g-r, r-i, i-z, log(f_X/f_R), y-j, j-h, h-k$	92.15%	93.18%
$hr1, hr2, hr3, hr4, i, u-g, g-r, r-i, i-z, log(f_X/f_R), y-j, j-h, h-k$	91.57%	92.78%

Table 2. The detailed accuracy for the best classification result by rotation forests

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
GALAXY	0.866	0.032	0.909	0.866	0.887	0.971
QSO	0.968	0.142	0.943	0.968	0.956	0.971
STAR	0.712	0.000	0.974	0.712	0.822	0.965

The accuracy with different input patterns is listed in Table 1. The detailed accuracy for the best classification result by rotation forests is shown in Table 2.

4. Conclusions

We compare the performance of random forests and rotation forests on the classification of all types of objects. Obviously, rotation forests are superior to random forests. The accuracy of classification improves when infrared information added, especially that of stars. Moreover all kinds of stars are easily discriminated and they have a clear distribution rule. In some parameter space, the special stars have different distribution from the most of stars. Therefore it is easy to choose special X-ray emitting star candidates from the huge photometric survey data depending on the information from optical and infrared bands. Firstly the rotation forest classifier is applied to select star candidates, then the special star candidates are picked out from the y-k vs. g-z diagram.

Acknowledgements

This paper is funded by National Key Basic Research Program of China 2014CB845700, National Natural Science Foundation of China under grants Nos.11178021, 11033001 and NSFC-Texas A&M University Joint Research Program No.11411120219. We acknowledge the SDSS, UKIDSS and XMM databases.

References

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. 2009, *SIGKDD Explorations*, 11(1), 10

Rodriguez, J. J., Escuela Politecnica Superior, Burgos Univ., Kuncheva, L. I., & Alonso, C. J. 2006, *Pattern Analysis and Machine Intelligence*, 28(10), 1619

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

https://doi.org/10.1017/S1743921315000022 Published online by Cambridge University Press