NATIONAL CENTRE FOR NUCLEAR RESEARCH

Abstract

Machine learning based catalogs of quasars and galaxies for cosmological studies

Szymon NAKONECZNY

We present two catalogs of quasars derived from the photometric Kilo-Degree Survey (KiDS) Data Release 3 and 4 (DR3, DR4). We present an approach to classification in KiDS DR3, and a more complex methodology to KiDS DR4, in which we add near-infrared imaging, estimate photometric redshifts, and extrapolate to magnitudes fainter than available in the spectroscopy. Our approach to KiDS DR4 produces the final quasar catalog, which we describe below. We build the catalog by training machine learning (ML) models, using optical *ugri* and near-infrared $ZYJHK_s$ bands, on objects known from Sloan Digital Sky Survey (SDSS) spectroscopy. We define inference subsets from the 45 million objects of the KiDS photometric data limited to 9-band detections, based on a feature space built from magnitudes and their combinations. We show that projections of the high-dimensional feature space on two dimensions can be successfully used, instead of the standard color-color plots, to investigate the photometric estimations, compare them with spectroscopic data, and efficiently support the process of building a catalog. The model selection and fine-tuning employs two subsets of objects: those randomly selected and the faintest ones, which allowed us to properly fit the bias versus variance trade-off. We tested three ML models: random forest (RF), XGBoost (XGB), and artificial neural network (ANN). We find that XGB is the most robust and straightforward model for classification, while ANN performs the best for combined classification and redshift. The ANN inference results are tested using number counts, Gaia parallaxes, and other quasar catalogs that are external to the training set. Based on these tests, we derived the minimum classification probability for quasar candidates which provides the best purity versus completeness trade-off: $p(\mathrm{QSO_{cand}}) > 0.9$ for $r < 22$ and $p(\mathrm{QSO_{cand}}) > 0.98$ for $22 < r < 23.5$. We find 158,000 quasar candidates in the safe inference subset ($r < 22$) and an additional 185,000 candidates in the reliable extrapolation regime ($22 < r < 23.5$). Test-data purity equals 97% and completeness is 94%; the latter drops by 3% in the extrapolation to data fainter by one magnitude than the training set. The photometric redshifts were derived with ANN and modeled with Gaussian uncertainties. The test-data redshift error (mean and scatter) equals 0.009 ± 0.12 in the safe subset and -0.0004 ± 0.19 in the extrapolation, averaged over a redshift range of $0.14 < z < 3.63$ (first and 99th percentiles). Our success of the extrapolation challenges the way that models are optimized and applied at the faint data end. The resulting catalog is ready for cosmology and active galactic nucleus (AGN) studies, and we perform an early study on constraining the quasar bias function, using its cross-correlation with the CMB lensing. We obtain a bias function $b_q(z) = 0.57^{+0.03}_{-0.03}(1+z)^2 + 0.07^{+0.06}_{-0.13}$, which at redshift $z = 1.5$ gives a bias value $3.63^{+0.25}_{-0.85}$. Finally, we report 15σ significance of the cross-correlation with the CMB lensing, which for quasars is one of the highest detections of this signal. We publicly release the catalogs at
http://kids.strw.leidenuniv.nl/DR4/quasarcatalog.php (KiDS DR4),
http://kids.strw.leidenuniv.nl/DR3/quasarcatalog.php (KiDS DR3).