XMM-Newton : a pathfinder for future multi-wavelength and multi-messenger observations with Athena

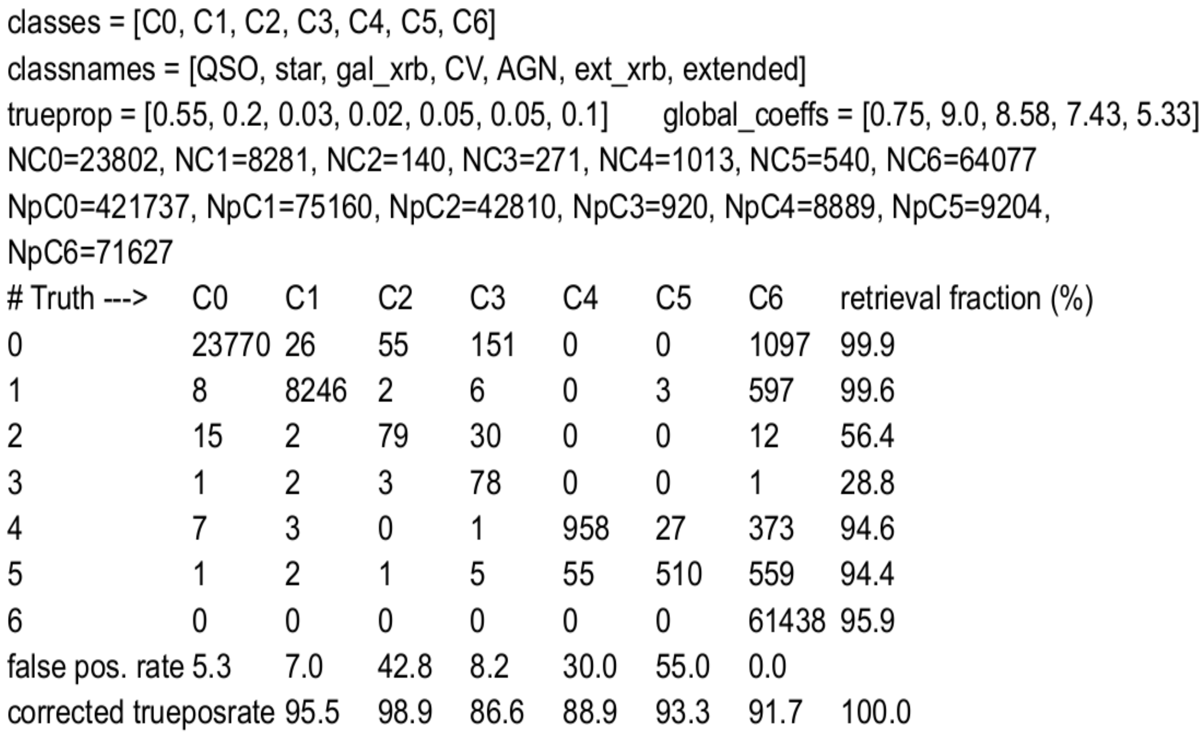
This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 101004168.

Work Package WP8
X-ray source classification

Deliverable D8.2 and D8.3
Implementation of classification AI algorithms report
Classification catalogue for 4XMM sources

Due date:	31.03.2023 and 30.09.2023
Nature[1]:	P
Dissemination Level[2]:	PU
Work Package:	8
Lead Beneficiary:	CNRS/IRAP
Contributing Beneficiaries:	UC, NOA, ULEIC, UCL, MPG
Version:	1.0

Contents



```
classes = [C0, C1, C2, C3, C4, C5, C6]
classnames = [QSO, star, gal_xrb, CV, AGN, ext_xrb, extended]
trueprop = [0.55, 0.2, 0.03, 0.02, 0.05, 0.05, 0.1]      global_coeffs = [0.75, 9.0, 8.58, 7.43, 5.33]
NC0=23802, NC1=8281, NC2=140, NC3=271, NC4=1013, NC5=540, NC6=64077
NpC0=421737, NpC1=75160, NpC2=42810, NpC3=920, NpC4=8889, NpC5=9204,
NpC6=71627
# Truth --->    C0     C1     C2     C3     C4     C5     C6     retrieval fraction (%)
0               23770  26     55     151    0      0      1097   99.9
1               8      8246   2      6      0      3      597    99.6
2               15     2      79     30     0      0      12     56.4
3               1      2      3      78     0      0      1      28.8
4               7      3      0      1      958    27     373    94.6
5               1      2      1      5      55     510    559    94.4
6               0      0      0      0      0      0      61438  95.9
false pos. rate 5.3    7.0    42.8   8.2    30.0   55.0   0.0
corrected trueposrate  95.5   98.9   86.6   88.9   93.3   91.7   100.0
```

Figure 1: The results of testing the classification of the catalogue. The 7 classes of objects (classnames) are shown at the top of the table, with the proportion of the sample for each class (trueprop) and the number in each class (NCx). The total population is also shown for each class (NpCx). The table shows the retrieval fraction for each class and the false and corrected true positive rates for each class (see also Tranin et al. 2022). The retrieval and true positive rates are high, indicating reliable results. X-ray binaries and CVs are difficult to retrieve, but the retrieved sample is fairly pure (>86% true positives).

1 Introduction

Following the classification work using a newly developed Naive Bayes classifier, presented in Tranin et al. (2022), we have further developed the algorithm and then applied the algorithm to each of the 630347 individual sources in the latest version of the *XMM-Newton* catalogue, version 4XMM-DR12 (Webb et al. 2020). We have extended the algorithm to identify Active Galactic Nuclei (AGN) alone or in a galaxy, extra-galactic and galactic X-ray binaires (XRB), stars, cataclysmic variables (CVs) and extended sources. This allows a source to be identified as Galactic or extragalactic. Following the classification of each source, we have generated a catalogue containing each of the 630347 sources from 4XMM-DR12 and augmented it with data from other multi-wavelength catalogues and provided a classification of each source. The catalogues are provided on the XMM2ATHENA webpages `http://xmm-ssc.irap.omp.eu/xmm2athena/` .

2 Validation

We have validated the reliability of this new classification through manual screening and testing with sources of known types. The results of the testing can be seen in Figure 1.

For further discussion on the merits of the classification, the reader is referred to Tranin et al. (2022).

3 Catalogue

The catalogue of identifications is provided as either a FITS (Flexible Image Transport System) file or as a CSV (Comma Separated Variable) file. These files provide one line per each of the 630347 sources. A number of columns from the 4XMM-DR12 catalogue (see `http://xmmssc.irap.omp.eu/Catalogue/4XMM-DR12/4XMM-DR12_Catalogue_User_Guide.html#Catalogue` for details on the columns) are also provided, including the source identification (SRCID), Right Ascension in degrees (RA), declination in degrees (DEC), position error in arcseconds (SC_POSERR), the four hardness ratios (SC_HRx) and their errors (SC_HRx_ERR), the source extent in arcseconds (SC_EXTENT) and the associated error, SC_EXT_ERR and the maximum likelihood of this extension (SC_EXT_ML).

In addition, the coordinates of the optical counterpart are provided (RA_Opt and DEC_Opt) and the distance between the X-ray source and the optical counterpart in arcseconds (angdist_Opt). The B and R band magnitudes are also provided (Bmag and Rmag respectively, equivalent to Gaia BP and RP filters) and the probability of the counterpart being the correct counterpart as calculated by NWAY (Salvato et al. 2018) (p_single_Opt and p_any_Opt). The catalogue from which the optical counterpart was derived is also provided (Ref_Opt), namely Gaia, PanSTARRS, USNO, etc. Similar information is provided for the infra-red counterpart i.e. RA_IR, DEC_IR, angdist_IR (arcseconds), W1mag, W2mag, p_single_IR and p_any_IR and the catalogue from which they are derived (Ref_IR). The Galactic longitude (l) and latitude (b) are provided and the Glade RA and dec, with the emi-major axis of the ellipse reprensenting the area of the associated galaxy (R1) and the semi-minor axis (R2), from GLADE 2016, the position angle (PA) and Distance to the galaxy (Dist) and the galaxy type (spiral, elliptical), with their associated probabilities (prob_sp and prob_el respectively). The separation of the Glade counterpart and the galactocentric distance are also given (Separation_GLADE and SepToRadius). The X-ray luminosity (Lx) (0.2-12 keV) and the log of the ratio of the fluxes between the X-ray and the b-band (logFxFb) and for the r-band (logFxFr) and for the W1 band (logFxFw1) and W2 band (logFxFw2) are also given. The Gaia proper motion (GAIA_pm) in miliarcseconds/year and Gaia distance (GAIA_Dist) in parsecs are given. A second X-ray luminosity is given Lx_2 (0.2-12.0 keV) calculated using the Gaia EDR3 distance.

The predicted nature of the source is given for AGN (isAGN), star (isStar), X-ray binary (isXRB) where 0 = no and 1 = yes. The output class given by the classification (0-6, where 0 = AGN, 1 = star, 2 = galactic X-ray binary, 3 = CV, 4 = background AGN, 5 = extended X-ray binary, 6 = extended source). The MASTER_ID gives unique identifier of the source across X-ray catalogues LSXPS, CSC2 and 4XMM-DR12, used to compute the variability between multi-mission detections. The flux ratio between the maximum and minimum fluxes, across all multi-mission detections of the source is given by fratio and the ratio between the maximum (flux-error) and minimum (flux+error), accross all multi-mission detections of the source. The proposed source type is given in prediction_name and this is also given as the class in the column prediction. alt gives alternative classifications if a property category is ignored, ClMargin gives the classification margin, i.e. P(prediction)-P(not(prediction)), outlier gives the outlier measure (see Tranin et al. 2022), N_missing gives the number of fields having a missing value, so that the reliability of the classification can be assess, PbaC0 gives the posterior probability that the source is an AGN, PbaC1 gives the posterior probability that the source is a star, PbaC2 gives the posterior probability that the source is a galactic X-ray binary, PbaC3 gives the posterior probability that the source is a CV, PbaC4 gives the posterior probability that the source is a background AGN, PbaC5 gives the posterior probability that the source is an extended X-ray binary and PbaC6 gives the posterior probability that the source is extended. PbaC0_location gives the combined likelihood of location properties for the class AGN, PbaC1_location, gives the combined likelihood of location properties for the class star, PbaC2_location the combined likelihood of location properties for the class galactic X-ray binary, PbaC3_location gives the combined likelihood of location properties for the class CV, PbaC4_location gives the combined likelihood of location properties for the class background AGN, PbaC5_location gives the combined likelihood of location properties for the class extended X-ray binary and PbaC6_location gives the combined likelihood of location properties for the class extended.

PbaC0_spectrum, PbaC1_spectrum etc are the combined likelihood of spectrum properties for the class AGN, star, etc. PbaC0_multiwavelength, PbaC1_multiwavelength are the combined likelihood of multi-wavelength properties for the class AGN, star etc. PbaC0_variability, PbaC1_variability, etc are the combined likelihood of variability properties for the class AGN, star etc. Small values ($< 10^{-4}$) imply that the source is likely to be variable.

3.1 Documentation

The full description of the classification method is found in Tranin et al. (2022). The column descriptions are given above, but also as meta-data in the catalogues, where the units and the class of the entry is also provided.

4 References

- Salvato M., Buchner J., Budavári T., Dwelly T., Merloni A., Brusa M., Rau A., et al., 2018, MNRAS, 473, 4937. doi:10.1093/mnras/stx2651
- Tranin, H., Godet, O., Webb, N., & Primorac, D., 2022, A&A, 657, A138, 10.1051/0004-6361/202141259
- Webb N. A., Coriat M., Traulsen I., Ballet J., Motch C., Carrera F. J., Koliopanos F., et al., 2020, A&A, 641, A136. doi:10.1051/0004-6361/201937353