*SF2A 2021*
*A. Siebert, K. Baillié, E. Lagadec, N. Lagarde, J. Malzac, J.-B. Marquette, M. N'Diaye, J. Richard, O. Venot (eds)*

# MINING THE HIGH-ENERGY UNIVERSE: A PROBABILISTIC, INTERPRETABLE CLASSIFICATION OF X-RAY SOURCES FOR LARGE X-RAY SURVEYS

H. Tranin[1], N. Webb[1] and O. Godet[1]

**Abstract.** Serendipitous X-ray surveys have been proven to be an efficient way to find rare objects – tidal disruption events, galaxy clusters, binary quasars, etc. As X-ray astronomy slowly enters the era of Big Data, an automated classification of X-ray sources becomes increasingly valuable. I present a revisited Naive Bayes Classification of the X-ray sources in the *Swift*-XRT and *XMM-Newton* catalogues which amongst other objects identifies different types of AGN, stars and X-ray binaries – based on their spatial, spectral and variability properties at different timescales and their multiwavelength counterparts. An outlier measure is used to identify objects of other nature. I show the reliability of the method developed and demonstrate its suitability to data mining purposes. As an outlook, I introduce how the very small populations in some object classes can be enlarged using citizen science, with the development of a new platform designed for the classification of XMM sources by volunteers.

Keywords: catalogs, X-rays: general, X-rays: binaries, methods: statistical,

## 1 Introduction

Since its beginning in the 1960s, X-ray astronomy has known significant breathroughs pushing its limits in both sensitivity and angular resolution. To this day, sources detected by *Swift*-XRT, *XMM-Newton* and *Chandra* facilities are gathered in large X-ray catalogues totalling about 1 million X-ray sources, and most of them remain unstudied. This ever-growing number of sources illustrates how X-ray astronomy is progressively entering the era of Big Data. Nevertheless, an automatic, efficient and interpretable classification of X-ray sources adapted to these large surveys is still to be developed. Such a tool will be of great interest e.g. to perform data-mining studies in X-ray archives, to send an alert when observing a new rare and exotic object – changing-look AGN (LaMassa et al. 2015), ultraluminous and hyperluminous X-ray sources (e.g. Farrell et al. 2009), tidal disruption events (e.g. Lin et al. 2018)... and to enable population studies of such objects. Previous attempts to classify X-ray sources generally focused on small samples of a few thousand objects, using different classification techniques such as decision trees (Lin et al. 2012), random forest (Farrell et al. 2015) and exploring other machine learning methods (Arnason et al. 2020). They classify X-ray sources using their properties such as the location, X-ray hardness and spectral parameters, X-ray short-term and long-term variability and multiwavelength counterparts, but never all at the same time. While decision trees are easy to interpret but lack efficiency, machine learning methods are more accurate but often black-box. In this work – more detailed in Tranin et al. 2021 – we develop a probabilistic classifier for the *Swift* and *XMM-Newton* catalogues, 2SXPS (Evans et al. 2020) and 4XMM-DR10 (Webb et al. 2020), intended to reach a good trade-off between efficiency and interpretability, and taking advantage of all the previously mentioned source properties.

## 2 Method

In order to obtain optimal classification results, we first enriched the X-ray catalogues with additional data:

- We identified the best optical and infrared counterparts for each source, using the bayesian crossmatching algorithm Nway (Salvato et al. 2018) and catalogues of optical and infrared sources – among other Gaia EDR3 (Gaia Collaboration et al. 2021) and UnWISE (Schlafly et al. 2019). This enabled the computation of the X-ray to optical (resp. infrared) flux ratios.

[1] IRAP, Université de Toulouse, CNRS, CNES, 9 avenue du Colonel Roche, 31028 Toulouse, France

- We computed the long-term variability as the ratio between the maximum and the minimum flux gathering all X-ray detections among *Swift*, *XMM-Newton* and *Chandra* observations (Evans et al. 2010).
- We identified the galaxies potentially hosting the X-ray sources using a positional cross-correlation with the GLADE catalogue (Dalya et al. 2016), containing 2 billion galaxies, their distance and angular size and rather complete up to 300Mpc.
- We identified known AGN, stars, X-ray binaries (XRB) and cataclysmic variables (CV) by a positional cross-correlation with catalogues covering these types (notably Véron-Cetty & Véron 2010; Secrest et al. 2015; Kharchenko & Roeser 2009; Liu et al. 2006, 2007; Mineo et al. 2012; Ritter & Kolb 2015.

Last but not least, a sample of sources of sufficient quality was selected following these criteria: having at least one reasonable detection according to the catalogue quality flags; and having at least two of these qualities: 1) an optical counterpart, 2) an infrared counterpart, 3) a signal-to-noise ratio greater than 10 or an acquired spectrum and 4) several X-ray detections. This resulted in a sample representing ∼65% of each X-ray catalogue, e.g. about 138000 and 371000 sources for 2SXPS and 4XMM-DR10, respectively. In each catalogue, approximately 19000 AGN, 5000 stars, 500 X-ray binaries and 300 CV are previously identified sources, and they constitute the training sample. The rest constitutes the test sample to classify.

| Category | Properties | $\alpha_t$ |
|---|---|---|
| Location | Galactic latitude, Gaia proper motion, Offset of the source to the host galaxy nucleus, X-ray luminosity from the host galaxy distance | 7.5 |
| Hardness | Hardness ratios, Exponent of the powerlaw spectral fit | 3.2 |
| Variability | $F_{\mathrm{X,max}}/F_{\mathrm{X,min}}$ in a single observation, $F_{\mathrm{X,max}}/F_{\mathrm{X,min}}$ between all observations | 6.0 |
| Flux ratios | (Optical) $F_X/F_b$, $F_X/F_r$, (Infrared) $F_X/F_{W1}$, $F_X/F_{W2}$ | 2.0 |

**Table 1.** Source properties used in the classification. $\alpha_t$ is the weighting coefficient of the category, fine-tuned to maximize classification performance.

The method we developed uses the distributions of about 15 source properties (detailed in Table 1 and split in four property categories) as probability densities to infer the source class. The distribution of each property and each class was modelled using a kernel density estimation (Sheather 2004) on the training sample. The result is illustrated in Figure 1 for two properties. Different properties characterizing the same category are combined by multiplying the likelihoods. The probability of a class $C$ given the source properties is a weighted product of the likelihoods inferred from each property category, $L(t|C)$, multiplied by a prior representing the prior proportion of sources of class $C$ (we used 66%, 25%, 7% and 2% for AGN, star, XRB and CV, respectively):

$$\mathbb{P}(C|data) = \frac{\mathcal{P}(C) \times \left(\prod_{t\in\{\mathrm{categories}\}} \mathcal{L}(t|C)^{\alpha_t}\right)^{1/\sum_{t\in\{\mathrm{categories}\}}\alpha_t}}{\sum_{C\in\{\mathrm{classes}\}} \mathcal{P}(C) \times \left(\prod_{t\in\{\mathrm{categories}\}} \mathcal{L}(t|C)^{\alpha_t}\right)^{1/\sum_{t\in\{\mathrm{categories}\}}\alpha_t}} \quad (1)$$

where the coefficients $\alpha_t$ of each category were fine-tuned to optimize the $f_1$-score ($f_1 = 2/(1/recall + 1/accuracy)$) of the classification on one chosen class, which was XRB in our study. This fine-tuning was performed by a differential evolution algorithm (Storn & Price 1997) and the coefficients converged towards the values shown in Table 1. The location and variability information are thus the most discriminant to identify XRB. Following equation (1), our method is thus a revised version of the Naive Bayes Classifier (Murphy et al. 2006), allowing to compute the probabilities of each class and directly relate them to the values of the source properties. On top of that, the numerator of this equation depends on the frequency of sources at the same point of the parameter space, so we used it as an outlier measure (O.M.) which allows us to spot objects with exotic properties. We then evaluated the performance of the classifier by analyzing the recall and accuracy of each class in both the training and the test samples.

## 3 Results

When applied to the training sample, the classifier returned the results detailed in Table 2. Overall, more than 97% of sources are correctly classified, with a particularly good performance on AGN and stars having $f_1$-scores higher than 0.98. The optimization on X-ray binaries led to great results for this class as well, while CV

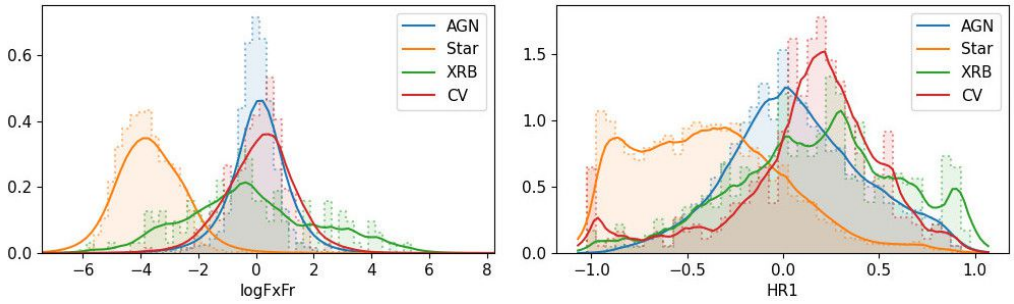


**Fig. 1.** Distributions of two properties in the reference sample of 2SXPS, and their kernel density estimation. **Left:** X-ray to r-band flux ratio. **Right:** Hardness ratio between soft and medium X-ray bands.

| | AGN | Star | XRB | CV |
|---|---|---|---|---|
| →AGN | 19515 | 82 | 25 | 191 |
| →Star | 44 | 4628 | 3 | 27 |
| →XRB | 140 | 18 | 326 | 17 |
| →CV | 9 | 9 | 2 | 124 |
| *recall* (%) | 99.0 | 97.7 | 91.6 | 34.5 |
| *precision* (%) | 97.0 | 98.6 | 90.7 | 85.5 |
| $f_1$-score | .980 | .981 | .911 | .492 |

| | AGN | Star | XRB | CV |
|---|---|---|---|---|
| →AGN | 18373 | 25 | 46 | 149 |
| →Star | 15 | 6197 | 10 | 12 |
| →XRB | 80 | 12 | 479 | 10 |
| →CV | 4 | 0 | 8 | 81 |
| *recall* (%) | 99.5 | 99.4 | 88.2 | 32.1 |
| *precision* (%) | 97.2 | 98.9 | 93.7 | 84.6 |
| $f_1$-score | .983 | .991 | .909 | .466 |

**Table 2.** Confusion matrices of the classifier applied to the 2SXPS (**left**) and the 4XMM-DR10 (**right**) training samples. The precision values are corrected for matching prior proportions.

are the most difficult to retrieve notably because of their diverse nature and the absence of detailed variability information in the enhanced catalogue. A detailed analysis of the XRB false positives revealed that most of them fall in one of these situations: they are an AGN in the background of a galaxy, a particularly variable AGN or star or their multi-instrument light-curve is not properly calibrated and thus shows a spurious variability. This diagnosis was easily obtained by looking at the sources' "probability tracks", a classification product showing the likelihoods of each class as given by each property (Figure 2).

According to the classification, the test sample is composed of about 80% of AGN, 17% of stars, 3% of XRB and 0.5% of CV. These proportions are in good agreement with the priors. A manual analysis of 200 sources revealed that more than 95% of AGN and stars were correctly classified, while sources classified as X-ray binaries contain about 50% of false positives because of the presence of objects of other nature and the reasons cited above. Enlarging the training sample is thus important for future progress, in order to refine XRB and CV types and represent rarer classes. Sources with a large outlier measure were also analyzed, showing a prevalence of spurious sources but also peculiar AGN and stars, XRB candidates, galaxy clusters and some transients.

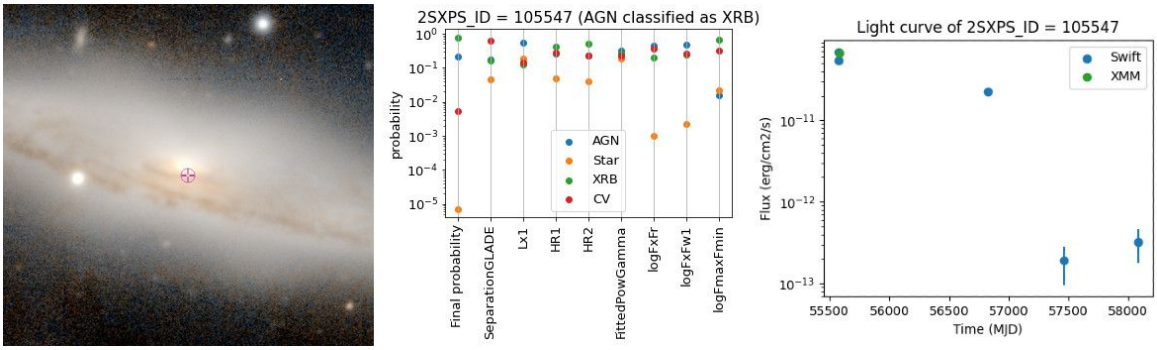


**Fig. 2.** 2SXPS J125801.1+013431, the central X-ray source of NGC 4845, known to host an AGN which underwent a tidal disruption event (Nikołajuk & Walter 2013). This source was wrongly classified as XRB. **Left:** PanSTARRS image of the galaxy and location of the source. **Middle:** Probability track of the source, showing the role of the variability ratio in the XRB probability. **Right:** X-ray light-curve from *Swift* and *XMM* detections.

## 4 Prospects

We developed a probabilistic, interpretable and efficient classification adapted to large X-ray surveys. After enhancing the *Swift* and *XMM-Newton* catalogues, we were able to classify more than 50% of their unknown sources. About 85-90% of these classifications proved to be reliable from a manual analysis, and each classification was made easy to interpret thanks to the classification products. Further research will address the applications of such a classification, and making a dynamic classification adapted to time-domain astronomy.

Besides, while AGN and stars are very well-classified, XRB and CV still show a lower performance which is at least partly due to their sparse and diverse training sample. In this context, enlarging the training samples e.g. using a citizen science approach is increasingly valuable. Citizen science takes advantage of the wisdom of crowds to ensure accurate classifications of samples as large as ∼100000 objects. Such experiments also proved to often lead to serendipitous discoveries. We thus launched CLAXSON (`http://xmm-ssc.irap.omp.eu/claxson`), a citizen science platform on which every volunteer can classify unknown X-ray sources after a discovery phase (quizz) and a training phase (classification of known sources and feedback). In order to classify a source, the user can examine its multiwavelength images and (when available) its spectrum and its short-term and long-term light curves. To this day, about 1000 unknown objects were classified more than 10 times thanks to 50 volunteers, who have a mean success rate of 82% in their classifications. Future work will therefore address the results of this experiment and evaluate its benefit in the field of X-ray classification.

This research has made use of several tools and services: Aladin sky atlas (`https://aladin.u-strasbg.fr/AladinLite/`) developed at CDS, Strasbourg Observatory, France (Bonnarel et al. 2000; Boch & Fernique 2014) ; TOPCAT version 4.8 (Taylor 2005).

## References

Arnason, R. M., Barmby, P., & Vulic, N. 2020, MNRAS, 492, 5075
Boch, T. & Fernique, P. 2014, in ADASS XXIII, ed. N. Manset & P. Forshay, Vol. 485, 277
Bonnarel, F., Fernique, P., Bienaymé, O., et al. 2000, A&AS, 143, 33
Dalya, G., Frei, Z., Galgoczi, G., Raffai, P., & de Souza, R. S. 2016, VizieR Online Data Catalog, VII/275
Evans, I. N., Primini, F. A., Glotfelty, K. J., et al. 2010, ApJS, 189, 37
Evans, P. A., Page, K. L., Osborne, J. P., et al. 2020, ApJS, 247, 54
Farrell, S. A., Murphy, T., & Lo, K. K. 2015, ApJ, 813, 28
Farrell, S. A., Webb, N. A., Barret, D., Godet, O., & Rodrigues, J. M. 2009, Nature, 460, 73
Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2021, A&A, 649, A1
Kharchenko, N. V. & Roeser, S. 2009, VizieR Online Data Catalog, I/280B
LaMassa, S. M., Cales, S., Moran, E. C., et al. 2015, ApJ, 800, 144
Lin, D., Strader, J., Carrasco, E. R., et al. 2018, Nature Astronomy, 2, 656
Lin, D., Webb, N. A., & Barret, D. 2012, ApJ, 756, 27
Liu, Q. Z., van Paradijs, J., & van den Heuvel, E. P. J. 2006, A&A, 455, 1165
Liu, Q. Z., van Paradijs, J., & van den Heuvel, E. P. J. 2007, A&A, 469, 807
Mineo, S., Gilfanov, M., & Sunyaev, R. 2012, MNRAS, 419, 2095
Murphy, K. P. et al. 2006, University of British Columbia, 18, 60
Nikołajuk, M. & Walter, R. 2013, A&A, 552, A75
Ritter, H. & Kolb, U. 2015, Acta Polytechnica CTU Proceedings, 2, 21
Salvato, M., Buchner, J., Budavári, T., et al. 2018, MNRAS, 473, 4937
Schlafly, E. F., Meisner, A. M., & Green, G. M. 2019, ApJS, 240, 30
Secrest, N. J., Dudik, R. P., Dorland, B. N., et al. 2015, ApJS, 221, 12
Sheather, S. J. 2004, Statistical science, 588
Storn, R. & Price, K. 1997, Journal of global optimization, 11, 341
Taylor, M. B. 2005, in ADASS XIV, ed. P. Shopbell, M. Britton, & R. Ebert, Vol. 347, 29
Tranin, H., Godet, O., Webb, N., & Primorac, D. 2021, submitted to A&A
Véron-Cetty, M. P. & Véron, P. 2010, A&A, 518, A10
Webb, N. A., Coriat, M., Traulsen, I., et al. 2020, A&A, 641, A136