Mining the high-energy Universe: a probabilistic, interpretable classification of X-ray sources for large X-ray surveys — The power of CLAXBOI

Hugo TRANIN, Postdoc, ICCUB, University of Barcelona

28 Feb 2024
Outline

1) Data preparation
2) Classification and interpretation
3) Applications
X-ray catalogs grow larger and larger

<table>
<thead>
<tr>
<th>Observations period, Coverage</th>
<th>PSF, Median Sensitivity</th>
<th>Number of sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>XMM-Newton 4XMM-DR13 (Webb+2020)</td>
<td>2000-2022, 1328 deg²</td>
<td>6’’</td>
</tr>
<tr>
<td></td>
<td>2000-2014, 560 deg²</td>
<td>0.5’’ on-axis</td>
</tr>
<tr>
<td></td>
<td>2005-2018, 3790 deg²</td>
<td>6’’</td>
</tr>
<tr>
<td>Chandra CSC2 (Evans+2019)</td>
<td>2000-2014, 560 deg²</td>
<td>0.5’’ on-axis</td>
</tr>
<tr>
<td>Swift-XRT 2SXPS (Evans+2020)</td>
<td>2005-2018, 3790 deg²</td>
<td>6’’</td>
</tr>
</tbody>
</table>
Focus of this talk

Observations period, Coverage

2000-2022
1328 deg²

PSF, Median Sensitivity

6”
1e-14 erg/cm²/s

Number of sources

657k

XMM-Newton
4XMM-DR13 (Webb+2020)

→ Expected content: AGN, stars, XRB, CV, galaxy clusters…
How to find them? ⇒ automatic source classification
1) Data preparation

“Prepare for battle” — Gandalf
Preparing the dataset for classification

1) Identification of known sources

- X-ray samples
- Catalogs of AGN (e.g. Secrest+2015)
- Catalogs of stars (e.g. Kharchenko+2009)
- Catalogs of XRB & CV (e.g. Ritter+2014)

TOPCAT software (Taylor+2005)
Sky with errors

(Simplistic crossmatch)

Ex. training sample of 4XMM-DR10

<table>
<thead>
<tr>
<th></th>
<th>AGN</th>
<th>Star</th>
<th>XRB</th>
<th>CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>19,000</td>
<td>6,000</td>
<td>730</td>
<td>260</td>
<td></td>
</tr>
</tbody>
</table>

Tranin et al. A&A 2022
Preparing the dataset for classification

2) Identification of counterparts

X-ray samples \( \otimes \) optical / IR surveys (Gaia, 2MASS…)

\[ \text{high sky density} \rightarrow \text{probabilistic treatment} \]

⇒ Multiwavelength associations

Flux ratios

\[
\log FxFr = \log_{10} \left( \frac{F_X}{F_R \text{ (Gaia)}} \right)
\]

Tranin et al. A&A 2022
Preparing the dataset for classification

3) Distance estimate

- Gaia distances (Bailer-Jones+2021)
- GLADE (Dalya+2016)
  TOPCAT Sky Ellipses Match

⇒ source distance & luminosity

\[ L_X = 4\pi D^2 \times F_X \]

GLADE = all-sky highly complete galaxy catalog

>1M galaxies at D<500Mpc

X-ray samples

ULX candidates

Tranin et al. A&A 2022

M51 in X-ray (XMM)

M51 in optical (PanSTARRS)
Multiwavelength dataset ready for classification

<table>
<thead>
<tr>
<th></th>
<th>Name / Reference</th>
<th>in 4XMM-DR11</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>X-ray samples</strong></td>
<td>-</td>
<td>496k</td>
</tr>
<tr>
<td><strong>Optical sources</strong></td>
<td>Gaia EDR3, PanSTARRS, DES</td>
<td>310k</td>
</tr>
<tr>
<td><strong>Infrared sources</strong></td>
<td>2MASS, AllWISE, UnWISE</td>
<td>420k</td>
</tr>
<tr>
<td><strong>Matches with galaxies</strong></td>
<td>GLADE (Dalya+2016)</td>
<td>16k</td>
</tr>
<tr>
<td><strong>Identified AGN</strong></td>
<td>Véron-Cetty+2010, Secrest+2015, Simbad</td>
<td>44k</td>
</tr>
<tr>
<td><strong>Identified Stars</strong></td>
<td>ASCC (Kharchenko+2009)</td>
<td>8k</td>
</tr>
<tr>
<td><strong>Identified CV</strong></td>
<td>Downes+2006, Ritter+2014</td>
<td>243</td>
</tr>
</tbody>
</table>
# Features used by the classifier

<table>
<thead>
<tr>
<th>Name</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Galactic latitude</td>
<td>Location</td>
</tr>
<tr>
<td>Gaia proper motion</td>
<td>Location</td>
</tr>
<tr>
<td>Relative distance to the host center</td>
<td>Location</td>
</tr>
<tr>
<td>X-ray luminosity</td>
<td>Location</td>
</tr>
<tr>
<td>X-ray over optical (b,r) flux ratio</td>
<td>Counterparts</td>
</tr>
<tr>
<td>X-ray over infrared (W1,W2) flux ratio</td>
<td>Counterparts</td>
</tr>
<tr>
<td>X-ray max to min flux ratio</td>
<td>Variability</td>
</tr>
<tr>
<td>X-ray lower max to higher min flux ratio</td>
<td>Variability</td>
</tr>
<tr>
<td>X-ray hardness ratio HR1, HR2, HR3...</td>
<td>Hardness</td>
</tr>
<tr>
<td>Power law index fitted to X-ray spectrum</td>
<td>Hardness</td>
</tr>
</tbody>
</table>
Physical properties:

- $\log F_x F_r$ (counterpart)
- $\log F_{\text{max}} F_{\text{min}}$ (variability)
- $HR_1$ (spectrum)
- $b$ (location)
- $\text{sep}$ (location)
- $L_x$ (spectrum)
2) Probabilistic classification (CLAXBOI) and interpretation

“You’re a wizard, Harry” – Hagrid
Methods for automatic source classification

Before 2022, in X-ray astronomy:

- Decision tree (e.g. Lin+2012) → poor performance
- Random forest (e.g. Farrell+2015, Arnason+2020) → poor interpretability
- Other machine learning algorithm (nearest neighbors, naive Bayes…) (e.g. Pineau+2017, Arnason+2020)

CLAXBOI: probabilistic classification, good interpretability and reliability
Previous studies

Previously classified samples (before 2022)

Small! ~ $10^{3-4}$ sources instead of $10^6$ detected

- Only bright sources (e.g. Lin+2012)
- Only variable sources (e.g. Farrell+2015)
- Only specific fields (e.g. Arnason+2020)

CLAXBOI: classification of most of well-detected point-like sources
Naive Bayes Classifier (2 classes)

Possible criterion:
\[
\log\left(\frac{F_{X}}{F_{W1}}\right) < -1 \Rightarrow \text{star}
\]
else \(\Rightarrow \text{AGN}\)

... but overlap

Tranin et al. A&A 2022
Naive Bayes Classifier (2 classes)

If \( \frac{F_x}{F_{W1}} = 0.01 \)
\( \mathcal{L}_{b, \text{Star}} = 93\% \)
\( \mathcal{L}_{b, \text{AGN}} = 7\% \)

If \( b = 50^\circ \)
\( \mathcal{L}_{b, \text{AGN}} = 67\% \)
\( \mathcal{L}_{b, \text{Star}} = 33\% \)

\[ P(\text{AGN}|D) = \frac{P(\text{AGN}) \mathcal{L}(\text{AGN}|D)}{P(\text{AGN}) \mathcal{L}(\text{AGN}|D) + P(\text{Star}) \mathcal{L}(\text{Star}|D)} \]

Combine the 18 features \( \Rightarrow \) Naive Bayes classification
Maximising the classification performance

- Trade-off between recall and precision
- Optimization: fine-tuning the $\alpha_t$

$$P(c|data) = \frac{P(c) \times \left( \prod_{t \in \{\text{cat}\}} \mathcal{L}(t|c)^{\alpha_t} \right)^{1/\sum_{t \in \{\text{cat}\}} \alpha_t}}{\sum_{C \in \{\text{classes}\}} P(C) \times \left( \prod_{t \in \{\text{cat}\}} \mathcal{L}(t|C)^{\alpha_t} \right)^{1/\sum_{t \in \{\text{cat}\}} \alpha_t}}$$

One $\alpha_t$ per category of properties: $\alpha_{\text{location}}$, $\alpha_{\text{spectrum}}$, $\alpha_{\text{variability}}$, $\alpha_{\text{counterparts}}$

Optimized to maximize the $f_1$-score of XRB  \( f_1 = (\text{recall}^{-1} + \text{precision}^{-1})^{-1} \)
Results (Confusion matrix)

Tranin et al. A&A 2022

⇒ better results on XRB + better interpretability
Interpretation #1: Finding outliers

\[ O.M. = - \log \left( \frac{P(c) \times \prod_{t \in \{\text{cat}\}} \mathcal{L}(t|c)^\alpha_t / \sum_{t \in \{\text{cat}\}} \alpha_t}{\mathcal{L}(t|c)^\alpha_t / \sum_{t \in \{\text{cat}\}} \alpha_t} \right) \]

\sim \text{scarcity of the training sample at the location of the source in the parameter space}
Depends on the output class \( c \)
⇒ way to nuance the classification

Outliers = one of these:
- Spurious sources
- Spurious identifications
- If classified as star/AGN: special types of star/AGN
- If classified as XRB: rare & variable objects such as TDE, GRB, supernovae...

Tranin et al. A&A 2022
Interpretation #2: marginal probabilities

Sources are classified based on their location, spectrum, counterparts and variability ⇒ find the discriminant properties thanks to marginal probabilities

\[ P_{\text{AGN}} = 88\% \]

Source inspection:
- Hard source
- No optical c. found
- little data

Marginal proba:
- spec and loc suggests Galactic XRB
- other+prior suggest AGN

⇒ classification as AGN is explained
Interpretation #3: alternative classifications

Sources are classified based on their location, spectrum, counterparts, variability.

What if we ignore a category of properties? ⇒ Alternative classification

Ex. previous source: no alternative classification
this blended source: alternative classification without location = Galactic XRB

SRCID=202004502010101

10 arcsec

P_{extended} = 92%

XMM extent 42”
Blends 3 Chandra sources
No opt or IR counterpart
Low Galactic latitude b=1°

XMM-Newton
Chandra
DESI Legacy Survey
3) Applications

“This is a beautiful tool but it still needs an active brain to use it”
– Mara Salvato
## Classification of a whole catalog

- **4XMM-DR12 fully classified (XMM2ATHENA deliverables)**
- **Published in April 2023:**
  

### 7 classes

Priors: 0.55, 0.20, 0.03, 0.02, 0.05, 0.05, 0.10

<table>
<thead>
<tr>
<th>truth →</th>
<th>AGN</th>
<th>Star</th>
<th>gal_XRB</th>
<th>CV</th>
<th>AGN_2</th>
<th>ex_XRB</th>
<th>extended</th>
</tr>
</thead>
<tbody>
<tr>
<td>→AGN</td>
<td>23770</td>
<td>26</td>
<td>55</td>
<td>151</td>
<td>0</td>
<td>0</td>
<td>1097</td>
</tr>
<tr>
<td>→Star</td>
<td>8</td>
<td>8246</td>
<td>2</td>
<td>6</td>
<td>0</td>
<td>3</td>
<td>597</td>
</tr>
<tr>
<td>→gal_XRB</td>
<td>15</td>
<td>2</td>
<td>79</td>
<td>30</td>
<td>0</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>→CV</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>78</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>→AGN_2</td>
<td>7</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>958</td>
<td>27</td>
<td>313</td>
</tr>
<tr>
<td>→ex_XRB</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>5</td>
<td>55</td>
<td>510</td>
<td>559</td>
</tr>
<tr>
<td>→extended</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>61438</td>
</tr>
</tbody>
</table>

**recall (%)**
- AGN: 99.9
- Star: 99.6
- gal_XRB: 56.4
- CV: 28.8
- AGN_2: 94.6
- ex_XRB: 94.4
- extended: 95.9

**precision (%)**
- AGN: 95.5
- Star: 98.9
- gal_XRB: 86.6
- CV: 88.9
- AGN_2: 93.3
- ex_XRB: 91.7
- extended: 100
Classification of a whole catalog

- 4XMM-DR12 fully classified (XMM2ATHENA deliverables)
  Published in April 2023:
  http://xmm-ssc.irap.omp.eu/xmm2athena/catalogues/

- **Content**
  430,941 AGN
  75,160 stars
  42,810 Galactic XRB
  8,889 extragalactic XRB
  920 Cataclysmic Variables
  71,627 extended sources

Priors: 0.55, 0.20, 0.03, 0.02, 0.05, 0.05, 0.10

Beware of spurious sources + crowded regions
Specialisation of the classification

X-ray samples \( \otimes \) GLADE (44k sources)

Goal: properly identify ULX candidates

<table>
<thead>
<tr>
<th></th>
<th>AGN (background sources)</th>
<th>Soft source (foreground sources, SNR)</th>
<th>XRB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>95.2</td>
<td>50.9</td>
<td>89.7</td>
</tr>
<tr>
<td>Precision</td>
<td>95.8</td>
<td>68.9</td>
<td>80.4</td>
</tr>
</tbody>
</table>

Goal: properly identify ULX
Identifying ULX in nearby galaxies

- A lot of interlopers remain here if we trust the maximum probability
- We need a physical prior and compare it with $P_{\text{XRB}}$
- Selection criterion: $P_{\text{XRB}} > f_{\text{contaminant}}$, frequency of background AGN from logN-logS

For the full population study check Tranin et al 2024, A&A 681 A16
CLAXBOI is public, documented and accessible via github (updated this week):  https://github.com/htranin/classificationXray

Feel free to use it for your science cases and reach me in case of questions!
Complementarity with citizen science

- CLAXBOI includes data preparation and value-adding
- Fully probabilistic classification
- Well-behaved on catalog-sized samples
- Both reliable and interpretable
- Samples of known XRB, CV, TDE… are still small

⇒ to enlarge training samples and find anomalies, use citizen science.
⇒ Tomorrow’s talk on CLAXSON
Conclusion

- **CLAXBOI** is a versatile, open-source and straightforward code to make the most of one’s X-ray catalog

- It can be easily tuned to identify X-ray sources in both **general** (entire catalogs) and **specific** (population study) frameworks

- It has been **successfully applied to 4XMM-DR12** (DR14 coming soon) but also CSC2, 2SXPS

- It provides **highly interpretable classifications**, helping scientific exploitation

- Automatic and Human-based source classification are complementary → see tomorrow’s talk about CLAXSON **citizen science project**
THANKS FOR YOUR ATTENTION