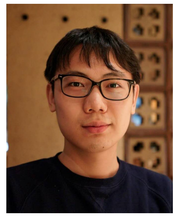
Classifying X-ray Sources with Supervised Machine Learning

Hui Yang[1] *, Yichao Lin[2], Oleg Kargaltsev[2], Steven Chen[2], Jeremy Hare[3]

[1] IRAP, Toulouse, France [2] Department of Physics, The George Washington University, Washington, DC, USA [3] NASA Goddard Space Flight Center,Greenbelt, MD, USA

*hui.yang@irap.omp.eu



Supervised Machine-Learning (ML) of X-ray sources

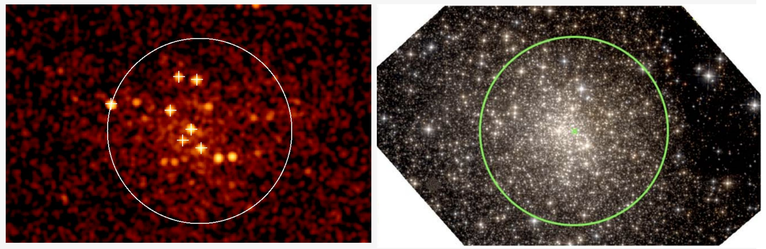


Fig. 1 Chandra and HST images of a globular cluster GLIMPSE-C01 field showing the same sky area. The solid white (left) and green (right) circles represent the 36″ half-light radius. Left: 180 ks Chandra image (0.5−7 keV). The white crosses indicate variable X-ray sources. Right: False-color HST image composed of F127M (blue), F139M (green), and F153M (red) WFC3/IR images.

- The X-ray universe is more dynamic and represents a different source population (e.g., isolated neutron stars, X-ray binaries), compared to the optical/IR universe.
- ML has been employed to classify X-ray sources detected by various observatories, including ROSAT (e.g., McGlynn+2004), Swift-XRT, XMM-Newton (e.g., Tranin+2022, Lin+in preparation), eROSITA (Salvato+2022), and our work (Yang+2022,2024) on Chandra Source Catalog version 2.0 (CSCv2).

The MUltiWavelength ML CLASSification Pipeline (MUWCLASS) on CSCv2 and 4XMM (Yang+2022, Lin+in preparation)

Training Dataset (TD)

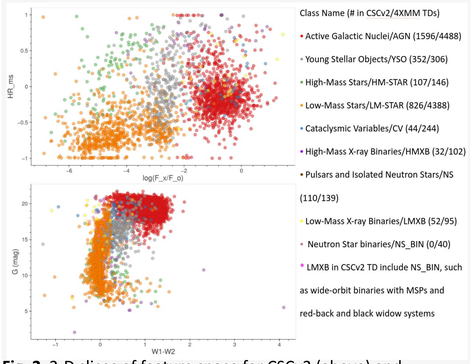


Fig. 2 2-D slices of feature space for CSCv2 (above) and 4XMM (below) TDs

Explore the TDs yourselves using the visualization GUI with QR codes here: CSCv2 TD (Yang+2021) 4XMM TD (Lin+2024) SCAN ME!

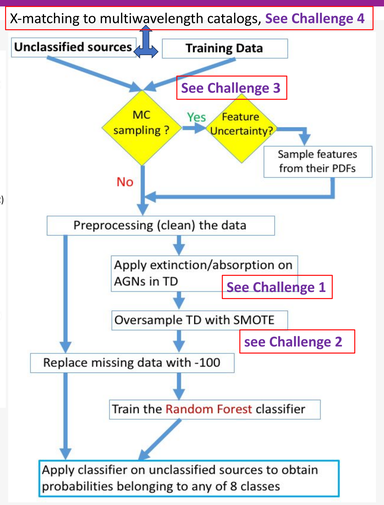


Fig. 3 The workflow chart of MUWCLASS pipeline

MUWCLASS performance on CSCv2 TD

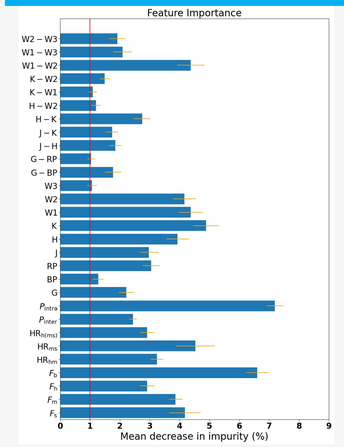


Fig. 4 The feature importance

- Feature importance = how often a particular feature is being used in the classification process.

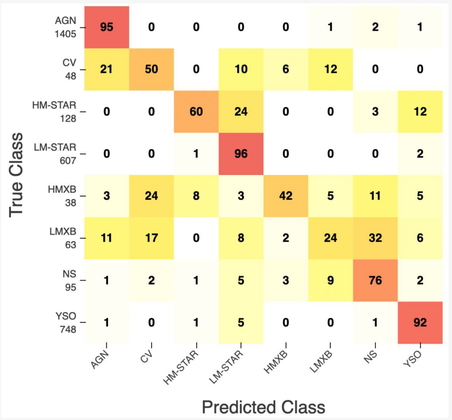


Fig. 5 Normalized confusion (**performance**) matrix

- A more diagonal matrix → better performance
- Not all classes are classified equally well.

Challenge 1: Biases between TD and unclassified sources

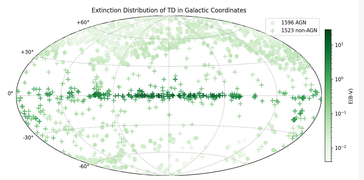


Fig. 6 The (SFD) E(B-V) dust map of TD sources in Galactic coordinates with AGNs shown in circles and non-AGN sources shown in crosses. A deeper color shows a higher value of E(B-V) (extinction).

- Most AGNs are situated off the Galactic plane, experiencing significantly less extinction than Galactic sources.
- An unclassified AGN within the Galactic plane looks much different than TD AGNs located off the plane.
- To address this bias, a direction-specific reddening (extinction and absorption) is applied to TD AGNs.

Challenge 2: Imbalanced TD

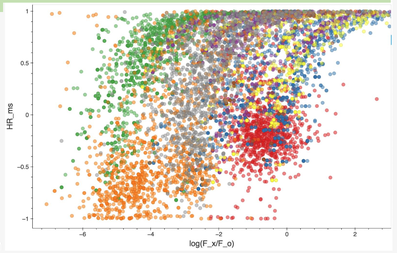


Fig. 7 "physically" oversampled TD for the same plot of Fig. 2.

- The TD is imbalanced (see the # of sources for each class in Fig. 2).
- We produce synthetic sources by sampling reddening parameters from those of TD and applying it on the less-populated (excluding AGN) class.
- This oversampling is more realistic/"physical" than other algorithms (e.g., SMOTE), and produce a fainter population of sources.

Challenge 3: Counting for Measurement Uncertainties

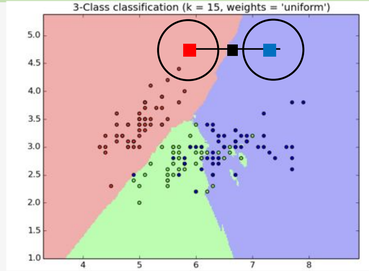


Fig. 8 An illustrative example of a 3-class classification in a 2-D feature space, showcasing how feature uncertainty (red and blue squares) of the source (black square) affects the classification.

- The impact of feature uncertainties is often underestimated (or ignored) in most ML works.
- We employ Monte-Carlo (MC) sampling to account for feature uncertainties by iteratively and randomly sampling feature values from their probability distribution functions.

Challenge 4: Cross-matching Confusion

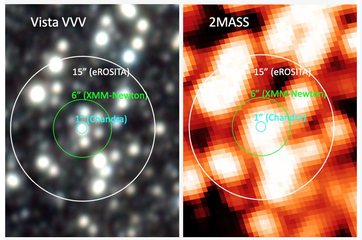
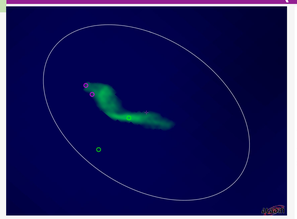


Fig. 9 The same sky region in VISTA VVV (left, a deeper survey) vs. 2MASS (right) overlapped with typical positional uncertainties (PUs) from Chandra, XMM-Newton and eROSITA.

- A probabilistic cross-matching method becomes essential when matching counterparts with deeper surveys (e.g., Pan-STARRs, DECaps, Vista VVV) of an X-ray source with larger PUs.

Exploring unidentified GeV sources with MUWCLASS (Yang+2024)



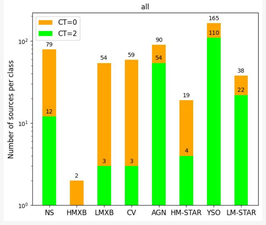


Fig. 10 The combined radio (RACS-low, in green) and TeV (HESS, In blue) image of 4FGL J1844-0306 while the white ellipse represents the 95% GeV error ellipse, along with the CSCv2 sources classified as NSs in magenta and YSOs in green. See more examples from QR code →

Fig. 11 The classification breakdown of CSCv2 sources within 73 GeV sources with green histograms marking the more confident classifications.

Summary

- We have developed an automated, multiwavelength machine learning pipeline, MUWCLASS, which has been applied to X-ray catalogs such as CSCv2 and 4XMM.
- We discuss common pitfalls often encountered in supervised ML, along with recent developments to potentially addressing these issues and improving MUWCLASS.
- MUWCLASS has also been used to identify particle accelerator candidates among unidentified GeV sources.
- Planned improvements include incorporating a 3-D extinction map, adding additional features (e.g., radio flux, distance), utilizing more sensitive surveys, and expanding TDs to Swift/XRT, and eROSITA catalogs.

References: • Mc Glynn, T. A., et al. 2004, ApJ, 616, 1284. • Tranin, H., et al. 2022, A&A, 657, A138. • Salvato, M, et al., 2022, A&A, 661, A3 • Yang, H., et al. 2022, ApJ, 941, 104. • Yang, H., et al. 2021, RNAAS, 5, 102 • Yang, H., et al. 2024, ApJ, 971, 180

Acknowledgement: This research is partly supported by NASA award 80NSSC22K1575 and by Chandra X-ray Observatory awards AR0-21007X and AR9-20005.