Machine-learning for real, messy data

University of BRISTOL

Sotiria Fotopoulou, Grant Stevens, Myank Singhal, Crispin Logan, Malcolm Bremer

Next generation telescopes will provide unprecedented amounts of data. At the same time, Cosmology experiments have very stringent classification requirements. The Euclid space mission aims to map the Dark Universe using Baryon Accoustic Oscilation and Weak grativational lensing. The later requires pure star and galaxies samples, as well as high quality photometric redshifts. ***We have created flexible machine-learning frameworks to deal with real, messy, astronomical data for the Euclid mission and beyond.***

Challenges

All photometric data show the following attributes:

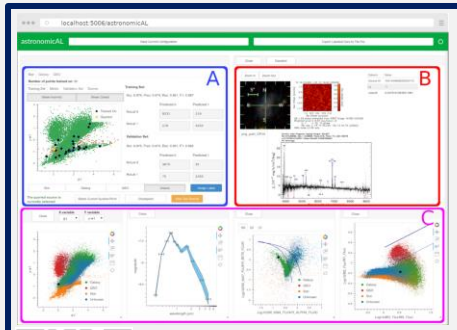
- Magnitude-dependent uncertainites
- Missing observations
- Potentially unreliable labels
- Labels originating from biased samples

Impactful scientific exploration requires **explainable, probabilistic classifiers** with rigorously assessed performance, including **uncertainty on their metrics**.

Domain knowledge and tailored use of machine-learning models leads to mitigation of these problems and to full exploitation of any additional information on the sources.

The work shown here is mainly based on photometry in the u-W1 bands presented in the CPz paper of Fotopoulou & Paltani (2018).

Tailored Solutions

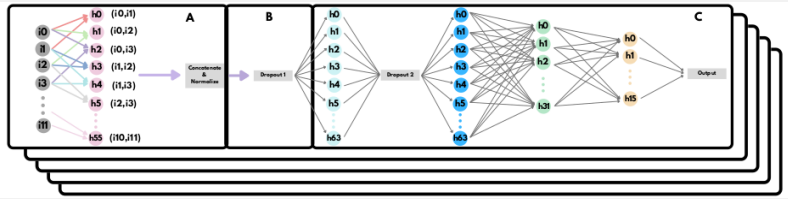


AstronomicAL
Stevens et al., 2021

AstronomicAL Active learning is a branch of supervised learning methods that traverses the training set with a smart approach. The model is initially trained with a handful of sources and the area of highest uncertainty is identified. The next source to be seen by the classifier is vetted by an expert, **mitigating unreliable labels**.

Stevens et al., 2021 presented **astronomicAL**, an active learning dashboard for tabulated data in astonomy. For more visit the QR code on the left.

StarMAP Neural networks offer the flexibility to work with missing & uncertain data.



Singhal et al., (in prep), created an **explainable** ensemble of fully connected networks which **tolerates missing data** during the training process.

Designed in particular for star-galaxy separation (F1=99%), this architecture takes into account measurement uncertainties, returns probability distribution functions for each classification, and provides uncertainty measurements on the performance metrics of the network.

Supervised or not?

HDBSCAN Unsupervised learning can be used to learn the structure of a high-dimensial parameter space, **bypassing the need for large labeled datasets**. Using clustering methods, similar objects are grouped together identifying dominant categories and enabling **outlier analysis**.

Logan & Fotopoulou (2020) showed that dimensionality reduction (PCA) combined with a density-based clusterer (HDBSCAN) leads to very high accuracy classification for stars (F1=98%), galaxies (F1=98.9%), and QSO (F1=93.1%).