Employing an Unsupervised Outlier Detection Algorithm to the Chandra Source Catalog

Dustin K. Swarm[1],
Casey T. DeRoo[1],
Randall K. Smith[2]

[1]University of Iowa, [2]Harvard-Smithsonian Center for Astrophysics

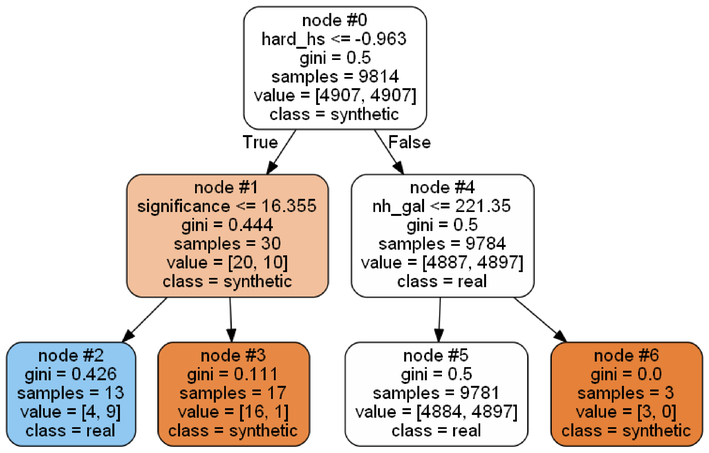

Figure 1: A sample decision tree trained to identify real and synthetic sources. Node color and shade indicates the class and node purity.

Unsupervised Random Forest

This work adapts the URF approach used by Baron & Poznanski 2017 to identify outliers in SDSS data. An RF is a collection of decision trees, with each tree in the forest trained on a random subset of features and training set sources. A URF is trained to differentiate between real and synthetic data with the same parameter distribution of the real sources but lacking feature covariance. The trained URF is reapplied to the real dataset. Sources with similar properties follow similar paths through the URF trees, allowing the URF to "categorize" the real sources.

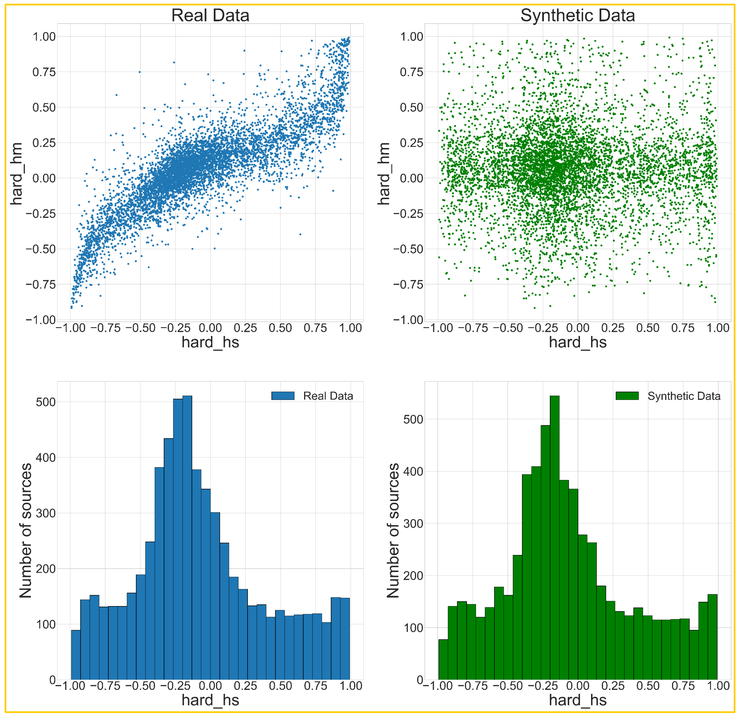

Figure 4: A comparison of real and synthetic data. (Above) A plot comparing the spectral colors. Synthetic data lacks the underlying covariance of real data. (Below) Compressing the above plots to one dimension shows the distribution of values in a feature are similar between real and synthetic sources.

Overview

The Chandra Source Catalog Release 2.0 (CSC 2) is a database comprised of approximately 317,000 unique X-ray sources observed by NASA's Chandra X-ray Observatory over the past two decades. We leverage this substantial data archive by employing a machine learning algorithm to search for unusual X-ray sources that have not been the subject of individual study. The algorithm applies an unsupervised random forest (URF) to the high-significance sources of the CSC 2, grouping together sources with similar observed properties. We compare results across many runs of the algorithm to find sources that are consistently grouped by themselves, showing they differ from other sources in the dataset. Looking for outlier sources within the CSC 2 increases the odds of discovering sources which offer new insight into astrophysical phenomena or are in a relatively short-lived evolutionary state.

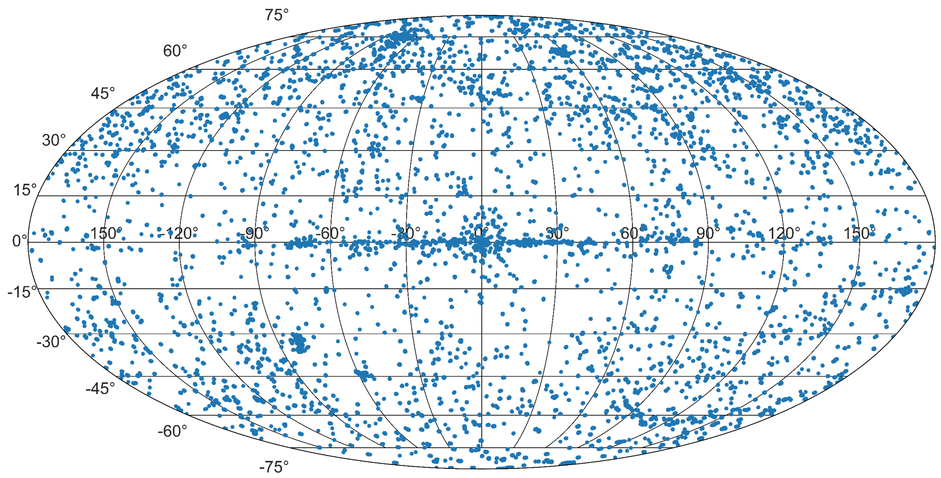

Figure 3: A galactic map of the ~35,000 CSC sources with significance greater than 7.5. Our analysis has focused on these sources to ensure identified outliers are reasonable targets for follow-up observations.

Outlier Sources

After applying the trained URF to the real sources, a "weirdness score" is calculated for each source by finding the average distance from that source to all others. Sources with a large average distance to other sources have a high weirdness score. Outlier sources are those with high weirdness scores across multiple applications of the URF algorithm using different initial random states and training sets.

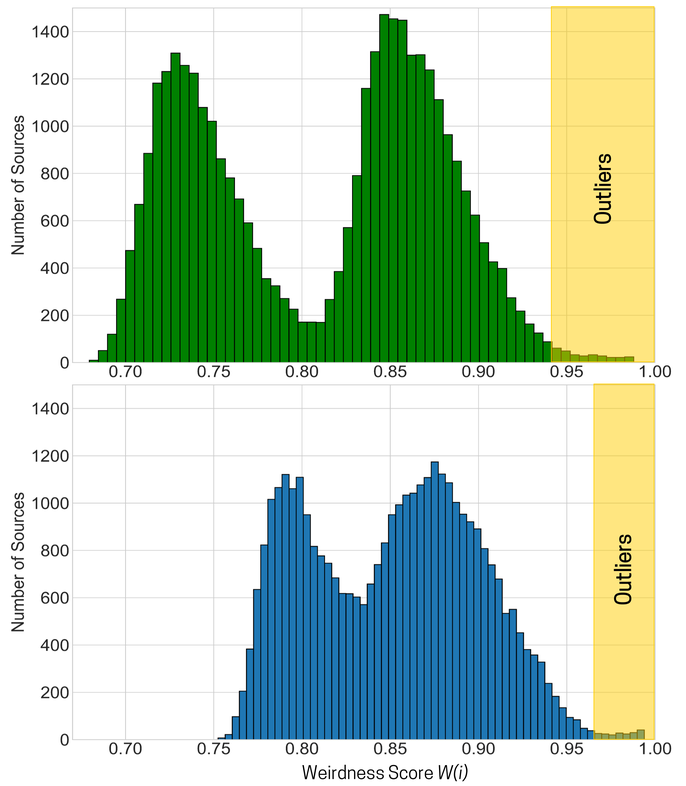

Figure 2: The distribution of W(i) for training sets of 1000 (green) and 6000 (blue) sources. The choice of hyperparameters affects the shape of the W(i) distribution as well as the repeatability of the algorithm. The 200 "weirdest" sources lie in the yellow shaded region on the right of each distribution.

Results and Future Work

Initial outlier searches have tracked the top 200 "weirdest" sources across multiple applications of the URF algorithm. This analysis produced 144 outlier sources that appeared in the top 200 weirdest sources across all runs of the algorithm. Many of these outliers are well observed but have not been specifically evaluated by astronomers. Follow-up spectral analysis will investigate the features that make these sources unique.

The specific choice of hyperparameters (e.g. the size of the training set or the number of decision trees in the random forest) alters the effectiveness of the URF algorithm. Further sampling of the hyperparameter space seeks to improve upon the algorithm's repeatability in this analysis process.

Acknowledgments

This project/material is based upon work supported by the Iowa Space Grant Consortium under NASA Award No. NNX16AL88H.

This work was supported by the University of Iowa College of Liberal Arts and Sciences.

References

Baron & Poznanski 2017. *MNRAS* 465 (4): 4530.