Iowa Initiative for Artificial Intelligence

Final Report

Project title:	Automated Searches of X-ray Data for the Earliest Black Holes
Principal Investigator:	Casey DeRoo
Prepared by (IIAI):	Yanan Liu
Other investigators:	Dustin Swarm (graduate student), Samantha Watkins (undergraduate)
Date:	11/05/2020
Were specific aims fulfilled:	Yes
Readiness for extramural proposal?	Yes
If yes … Planned submission date	1/14/20, 3/23/20
Funding agency	NASA, Chandra Science Center
Grant mechanism	Research Grants
If no … Why not? What went wrong?	

Brief summary of accomplished results:

With the assistance of the IIAI, we established a finalized catalog of outlier sources present within the Chandra Source Catalog, a collection of 300,000+ sources observed by the Chandra X-ray Observatory over two decades. The outlier catalog includes 119 sources, of which approximately 50% have not been previously studied.

Finalizing the catalog required a rigorous methodology study, for which IIAI assistance was invaluable. First, the astronomical data are not homogenous – approximately 50% are missing parameters derived from spectral analysis. Eliminating these sources a la listwise deletion is a non-starter, as early examinations of our sources showed that these spectral parameters feature prominently as criteria for the decision tree.

We identified a mode for proceeding to be able to analyze the sources missing this data (dummy variable adjustment), and quantified the classification bias present in this method of data cleaning by examining the number of synthetic vs. real sources misclassified. In addition, we performed a detailed metaparameter study to identify the number of sources per leaf, training set size, etc. needed to result in maximal repeatability from run to run. Accomplishing these goals has (1) set us up to publish papers on this catalog of sources that includes a rigorous technical description uncommon for machine learning work in astronomy and (2) will enable us to propose to agencies who are primarily concerned with the analysis of sources rather than the mechanics of machine learning.

Research report:

Aims (provided by PI):

At the outset, our goals were to identify black holes in the early Universe within the Chandra Source Catalog. However, additional examination revealed that the dataset did not have the sensitivity needed to find many of these sources in the first place, and that there was no “ground truth” catalog that we could employ as a training set.

Thus, our aims transitioned to identifying outliers that had been observed such that they were detected at high significance but had not been previously studied in-depth. This is consistent with new trends in astronomical research, which will require the identification of rare sources which improve our understanding of relevant physical processes out of millions of candidates. We were particularly focused on generating a robust, repeatable catalog of outlier sources (i.e. one not dependent on the particular training set) and understanding how to make defensible choices for a methods section of a proposal/paper.

Data:

The data are a collection of astronomical attributes (position in the sky, X-ray “color”, spectral fit parameters) for the 315,000 sources observed by the Chandra X-ray Observatory over its 20 year operation. The bulk of these sources are observed “serendipitously,” where sources are included in the field of view while Chandra is staring at an another object.

AI/ML Approach:

The outlier detection is based on unsupervised random forest (RF) since we don't have labeled data. We implement unsupervised RF with some modifications, in which we produce a synthetic dataset with identical parameter distributions as the real dataset, construct a random forest that can distinguish between real or synthetic, then resort the real dataset using the constructed random forest. Real sources that appear isolated via a distance measure are identified as outliers [1].

Experimental methods, validation approach:

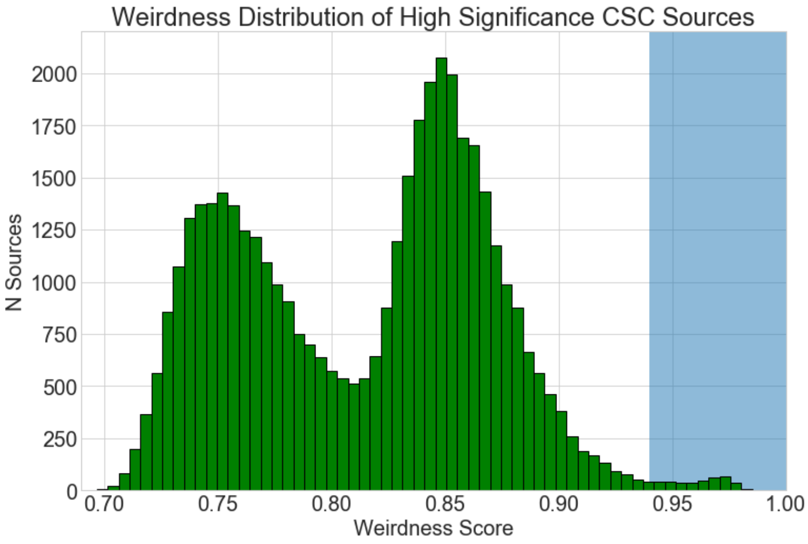
To validate our approach, we both conducted a bias study and a repeatability study in terms of metaparameters. For the metaparameters, we identified the training set size, selection criterion (Gini or entropy), number of sources per leaf, etc. that resulted in the same sources being identified as outliers from run-to-run. This was interpreted as having converged on these being outliers amongst the dataset regardless of the contents of the randomly-drawn training set. The chosen metric was total number of unique sources over 10 runs for a given set of metaparameters – fewer unique sources indicates greater agreement over the 10 runs on the set of outliers.

After establishing the metaparameters, we quantified the level of bias present in our decision trees. Sources were correctly classified 99.9% of the time by the decision tree, indicating that the construction of the tree was easily able to distinguish between synthetic and real sources and hence instilling confidence that the decision tree was constructed in such a way to isolate real outlier sources away from the vast majority of usual outliers.

Results:

See several example figures below.

Figure 1: Histogram of the "Weirdness" distribution of CSC sources as generated from the unsupervised RF algorithm. The weirdness score is a normalized measure of how often sources end up in different classification groups than the other sources in the set. Hence, a high weirdness score indicates that these are unlike other objects in the data set i.e., are outliers. The outlier sources in this histogram are identified by the blue transparent box.



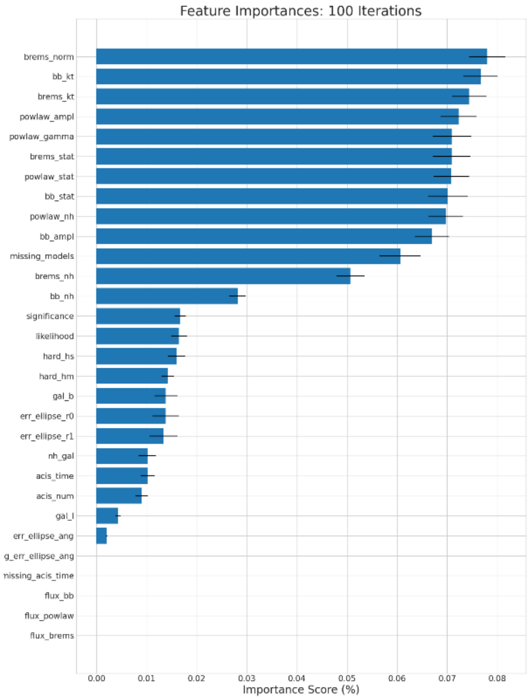


Figure 2: A chart showing the relative feature importance of the source parameters used in the random forest. This is normalized to unity, such that the importance score directly related to the percentage of decision points that use that parameter. Features with the greatest importance (brems_norm, kb_tt, etc.) are missing in about 50% of the sources due to low source counts. ***Our results have implications for both how ML algorithms should handle missing features in an astronomical case, as well as how astronomical catalogs should be reported/constructed in the first place.***

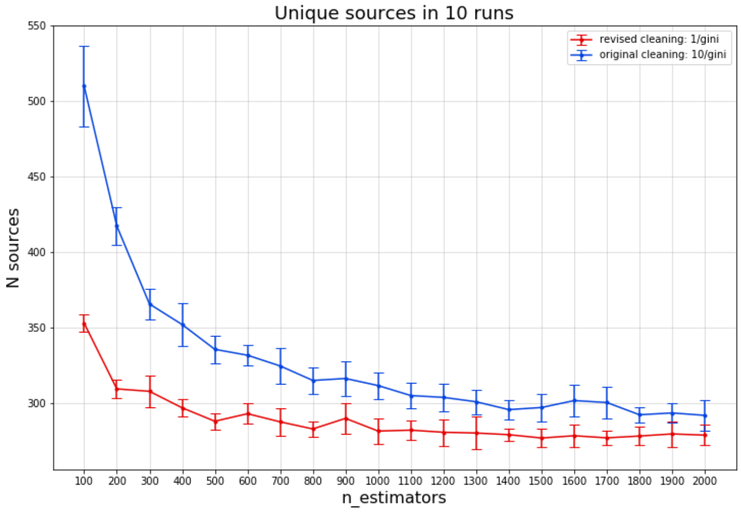


Figure 3: A diagnostic plot showing the effect of different algorithm parameters ("hyperparameters") on the number of unique sources identified in 10 runs of the RF. Fewer unique sources indicates better repeatability i.e., that the same outliers are found regardless of random starting values. ***IIAI personnel helped to identify the importance of these parameters as well as determine a method by which they could be selected, providing essential support for future proposals / publications.***

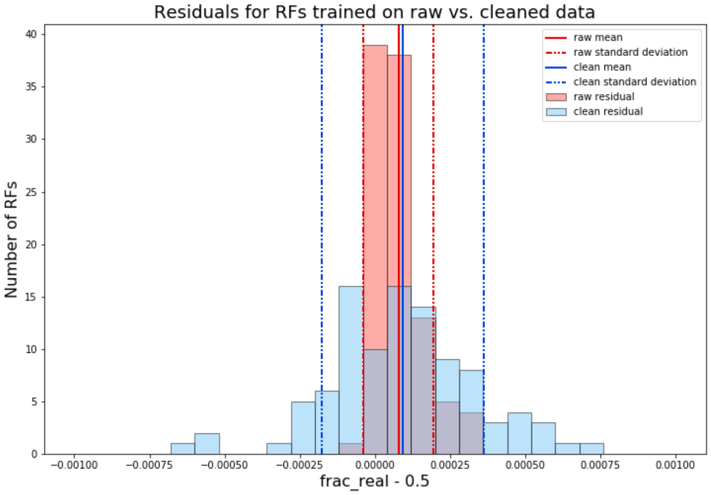


Figure 4: A diagnostic histogram assessing bias in the RF process. There is no consensus in astronomy on how to handle missing values from catalogs, which can include significant fractions of the overall data (~50%). ***With IIAI support, we were able to identify a method of cleaning our data for use with the unsupervised RF, dummy variable adjustment, while preserving the information present for some sources.*** *This histogram shows how often the RF decides a source is real given an input fraction of 50% real sources, and is thus a measure of bias. Using dummy variable adjustment (blue) results in a mild increase in the number of sources classified incorrectly (0.05%) as compared to the raw data (red), but shows no preference towards classifying a source as fake or real, validating this approach for the CSC dataset.*

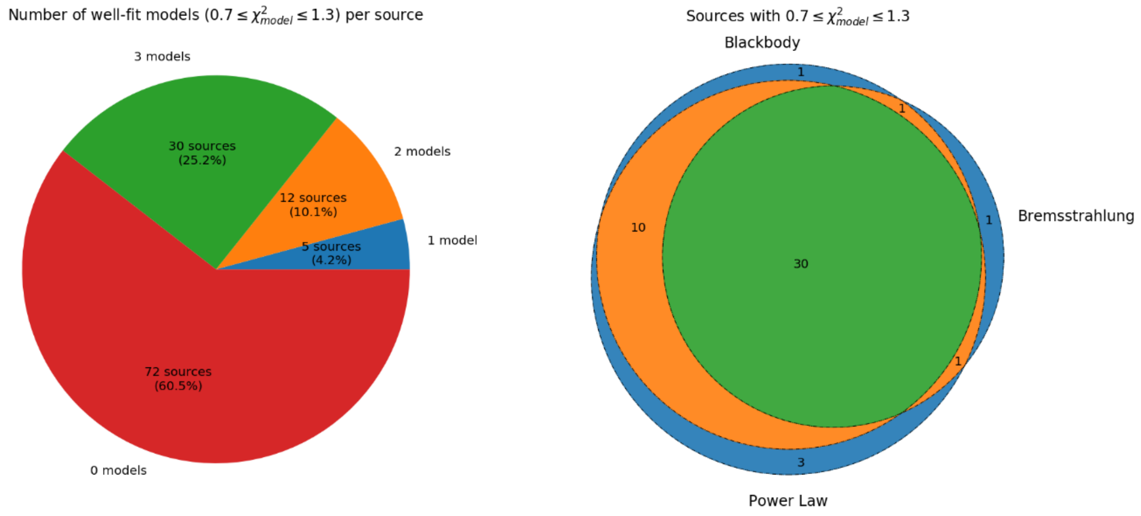


*Figure 5: Pie chart representation of the 119 outlier sources. Left: the number of sources best represented by a given spectral model. These models are mutually exclusive, meaning that the natures of 42 sources described by more than one model are unclear. Moreover, there are 72 sources not well-described by any of the standard models. Right: the distribution of models that could be used to describe the sources. Most of this subset of sources are described by a powerlaw, which does not yield physical insight into the nature of the source without further examination***. The vast majority of our outlier sources are not described by standard, well-understood physical models, making them promising targets for follow-up.***

Ideas/aims for future extramural project:

We are currently studying methods of data cleaning for astronomical application. As astronomical data is rarely homogenous and is subject to substantial selection bias (bright, nearby sources are observed while outlier sources may be dim, distant, or time-variable), concocting a set of recommendations for identifying unusual sources is of interest. We plan to employ sources from the Chandra Source Catalog that have all parameters present as the dataset, and identify the outliers there as the "ground truth." We will then make this dataset sparser by eliminating attributes in both random and structured scenarios, employ different treatments to this missing data (e.g., dummy variable adjustment, regression), and see which technique gets us outliers closest to the "ground truth" – the outliers found if you have access to all the data. This project is led by an undergraduate researcher, and will be the subject of her undergraduate thesis.

Publications resulting from project:

None yet, two planned (catalog publication and cleaning methodology study).

Bibliography:

[1] D. Baron and D. Poznanski, "The weirdest SDSS galaxies: Results from an outlier detection algorithm," *Mon. Not. R. Astron. Soc.*, vol. 465, no. 4, pp. 4530–4555, 2017.

External Proposal Success/Reviews:

- NASA EPSCoR R3 CAN (proposal # 20-EPSCoR2020-0040, "Using Machine Learning Techniques to Identify Unusual X-ray Sources), budget: $100,000
    - Proposals are intended to, among other things, contribute to and promote the development of research capability in NASA EPSCoR jurisdictions in areas of strategic importance to the NASA mission, and develop partnerships among NASA research assets, academic institutions, and industry.
    - We proposed to form a partnership with the Computational and Information Sciences and Technology Office (CISTO) located at NASA's Goddard Space Flight Center, leveraging some professional contacts of Prof. DeRoo's. **The work was found to have high intrinsic merit, suggesting that the proposal could be successful in the future**.

Reviewer Response:

Proposal #	Review Acronym	Proposal Jurisdiction	Proposal Title	Intrinsic Merit	Intrinsic Merit Strengths and / or Weaknesses Comments	Management, Coordination, and Evaluation
20-EPSCoR2020-0040	CISTO	IA	Using Machine Learning Techniques to Identify Unusual X-ray Sources	High probability for successful implementation	This is a high-quality proposal. The goals and objectives are clear, it addresses the expectations described in the announcement, and it demonstrates a high probability for successful implementation. The ML algorithms and procedures are well described, and methods have been chosen which have demonstrated successful applicability in the literature. The problem of unsupervised outlier detection is very interesting and has applications outside of astronomy.	Adequate detail provided
20-EPSCoR2020-0040	CISTO	IA	Using Machine Learning Techniques to Identify Unusual X-ray Sources	Probability for successful implementation uncertain	Investigators propose to adapt unsupervised object outlier algorithms (OIAs) used in other astronomical domains for use in X-Ray astronomy. Work would fundamentally involve a trade study, followed by algorithm adaptation based on the results of the trade study and consultation with NASA subject matter experts. Work has high intrinsic merit, but little detail is provide about how the work would be carried out, which makes it difficult to evaluate probability for successful implementation	Poor or not addressed

- Chandra X-ray Observatory (proposal # 22900269 "Identifying Unusual Sources in the Chandra Source Catalog"), budget: $85,000
    - Proposals are intended to support the analysis of archival Chandra X-ray Observatory data or (more frequently) support the analysis of proposed observations.
    - We proposed to develop a catalog of outlier sources from the Chandra Source Catalog using the ML techniques outlined above, and publish a paper of the identified sources for the communities' benefit.
    - The single reviewer did not find a blind search of the CSC compelling ("it is not clear what to expect from the scientific return of this proposal," which is in direct conflict with the idea of a blind search). **It is likely that we had a reviewer who has no machine learning background or expertise, and thus found our proposal hard to understand. In the future, we will propose for follow-up observations on outlier targets, which are funded at a higher rate than archival analysis proposals.**

Reviewer Response:

Review: Chandra Peer Review Form for 22900269

Proposal Number: 22900269

Subject Category: EXTRAGALACTIC DIFFUSE EMISSION AND SURVEYS

Joint: None

P.I. Name: Casey Thomas DeRoo

Proposal Title: Identifying Unusual Sources in the Chandra Source Catalog

---

Review Report:

Importance of Science = Good

Proposal Science Justification = Average

Feasibility = Good

Feasibility of Science if constraint preferences not met = N/A

Use of Chandra capability = Average

Clarity of proposal = Average

The proposal aims to identify unusual sources in the Chandra Source Catalog 2 using unsupervised outlier identification algorithm (UOIA). The algorithm is based on the combination of Random Forest and Unsupervised techniques widely used in machine learning. The goal is to build a catalog of unusual sources suitable for follow-up studies.

Strengths

- The algorithm has been successfully applied to 2 million SDSS sources and a rank-ordered estimate of the source "weirdness" parameter to prioritize targets for additional analysis has been created. As a proof-of-concept the proposers applied the unsupervised algorithm to a subset of the Chandra Source Catalog 2 identifying a preliminary list of unusual sources. The preliminary results look promising.

- Finding unusual objects is an important point at this time with the advent of surveys that will provide detection of billions of objects (e.g. Euclid, eRosita, LSST).

Weaknesses

- The structure, and highly technical nature of the proposal made it difficult to read.
- A discussion of any previous application of this or similar algorithms to Chandra catalogs would have been useful.
- It is not clear from the proposal how the hyper-parameters of algorithm will be optimized and the minimum size of the training set be varied in order to obtain a consistent and accurate result.
- In its current form, it is not clear what to expect from the scientific return of this proposal. Some list of sample science goals would be useful, even if of course there could be unexpected discoveries or surprises.

Degree of effort required to achieve analysis goals = Average
(flag used to adjust funding if proposal is approved)

- Grade: 2.97