In this application, I have included the abstract and summary of master's thesis. One manuscript based on the master's project thesis is under preparation, and therefore I would not be able to share my entire thesis as of now.

Probabilistic Classification of *Chandra* X-ray Sources Using Machine Learning

A thesis submitted for the award of the degree of

Master of Science

in **Astronomy and Astrophysics**

by

Shivam Kumaran

Supervisor:
Dr. Samir Mandal (IIST)
Dr. Sudip Bhattacharyya (TIFR)
Dr. Deepak Mishra (IIST)

Department of Earth and Space Sciences
Indian Institute of Space Science and Technology
Thiruvananthapuram, India

Abstract

Context : *Chandra* is an advanced X-ray telescope and have a coverage in 0.1-10keV energy band. It has detected a huge number of X-ray sources. With the all sky survey data from *Chandra*, Chandra Source Catalog (CSC-2.0) was prepared, that contains about 317,000 X-ray sources (296473 X-ray point sources). Out of these, the classification of a major fraction of sources remain unknown. We can not classify such a large set of sources by traditional methods that need an in-depth analysis for each individual source. Such classification on huge dataset can be done using automated classifier based on machine learning methods.

Aim: In this work, we aim to develop an automated machine learning-based model to classify the point sources in Chandra Source Catalog (CSC-2.0) and assign class membership probabilities to the sources.

Method : Out of the 296473 point sources, we filtered 277717 objects as good source based on quality flag in CSC. We did a literature survey to find the already identified sources that belongs to one of the classes in Active Galactic Nuclei, Young Stellar Objects(YSO), X-ray emitting star, High Mass X-ray Binary (HMXB), Low Mass X-ray Binary (LMXB), Cataclysmic Variables, Pulsar, and Ultra Luminous X-ray sources (ULX). Using the position cross-match with a radius of 1 arcsec, we identified the nature of 7703 sources in CSC and used them as the training set. We then found their counterpart in other wavelengths from Gaia, SDSS, GALEX for optical and UV, and 2MASS, WISE, and MIPS for Infra-red. We populated our data table from these multi-wavelength catalogs combined with the properties from CSC. We evaluated the classification performance of the network-based classifier and Decision tree-based ensemble classifier - Random Forest, Gradient Boosted decision tree, and Light Gradient Boosted Machine (LightGBM). We tried the classifier models in combination with different data imputation methods. With a custom version of K-fold cross validation, we found that LightGBM outperforms other classifiers and works with missing values, eliminating the need for data imputation.

Result : With the LightGBM classifier, we achieved 93% accuracy, 80% precision score, 79% recall, 80% f1 score, 0.905 Mathew's Correlation Coefficient score. To demonstrate the classifier scientific capabilities we applied it to the set of all the variable point sources in CSC-2.0 and identified the class of 15283 new variable sources with a probability of more than 0.98. We applied the classifier on globular cluster (GC) NGC-104, studied population distribution and identified 85 new Pulsars and 14 new CVs within a radius of 2.79 arcmin from the center of the GC.

Summary

Our aim in this work was to develop a model for the automated classification of sources in the Chandra Source Catalog. We prepared a list of 277717 good point sources from the CSC as our target of interest. Then we prepared the training set using a literature survey of sources belonging to our class of interest: AGN, YSO, X-ray emitting Stars, HMXB, LMXB, Pulsars, CVs, and ULXs. For all these sources we collected the flux in the various band of *Chandra* (broad band (b),ultrasoft(u), soft(s), medium (m) and hard(h)bands), variability properties in broadband, and hardness properties. We used Gaia-EDR3, SDSS and GALEX to obtain Optical/UV magnitudes, 2MASS, WISE and MIPS for infra-red magnitudes. To obtain data from 2MASS, WISE and MIPS we used NED catalog with CSC-2.0 names as identifier and for SDSS, Gaia-EDR3 and GALEX we used positional cross-match within 3-σ of positional error of the source.

We could identify the class of 7703 sources using the position cross-match with a radius of $1''$, from the literature survey. Then we identified the best data normalization method. After preparing the data table, we tried various machine learning classification models combined with different data imputation techniques. We tried imputation of missing values using feature-feature correlation, using statistical quantities and using ML iterative regression. We found out that for prediction of missing values, the regression imputation using ML regressor works best but it also removes the intrinsic difference in the properties across the class and is not a good choice for any classification task. Hence for classification imputation of missing values is not recommended.

On the final dataset we tried classifier based on neural network (NN) and various other classifiers based on Decision Tree (DT). Since we have features in tabular form and we do not need to do feature extraction DT based method outperformed NN based models. The Decision tree (DT) based ensemble models that we tried are: Random forest, Gradient boosted Decision Tree, and LightGBM. With NN based model we could achieve a maximum accuracy of 80%. Out of DT based model, LightGBM outperformed others. One other significant advantage we saw for LightGBM is that we can use it even with missing values in our data table, and we do not need to do any imputation. Neural network based classifier could achieve only 80% accuracy. With LightGBM we got the highest accuracy of 93% as compared to RF and GB that gave an accuracy of 89% .We fur-

ther tuned the hyperparameters of the LightGBM model and saved the trained model. LightGBM classifier gave precision 80%, recall 79% and MCC score of 0.905.

We developed a classification pipeline where for a given list of sources, the pipeline outputs class membership probabilities for the sources for each of the class in AGN, HMXB, STAR, YSO, LMXB, ULX, CV and PULSAR. To demonstrate the science capabilities we applied the trained model to the set of 37873 previously unclassified variable point sources in CSC-2.0 and identified the classification of 15283 of them with probabilistic confidence of more than 0.98. We also applied the model to GC NGC-104 sources. In NGC-104 we identified 141 Pulsars and 38 CVs with probabilistic confidence of more than 0.9. We then compared our classification result with that of the published work and found our result in a good agreement with them with only 6 sources out of our classification being of different class in the literature.

Future Work

At the current stage of our work, we have used only the observed features for the sources from various multiwavelength catalogs combined with the Chandra Source Catalog. With these properties using tuned LightGBM classifier we are able to classify AGN, STAR, YSO, HMXB and LMXB with a very high accuracy. For improving the score of other classes , CVs, ULX and Pulsars we would try the following ideas :

- Using feature from lightcurves for all the sources, such as the number of flares and periodicity would help us identify the Pulsars with high confidence.
- We would extract features from Spectral Energy Distribution, which can be further helpful to distinguish ULXs from Pulsars.

Currently we are dealing with small number of dataset for the CVs, LMXBs ULX and Pulsars. We will use advanced Deep Learning (DL) methods like Generative Adversarial Networks (GANs), which can generate more accurate synthetic data from the available dataset, thus we can further enhance our classifier performance.

Currently we have included only 8 classes in our model, but some of the sources may not belong to any of these classes. For such sources we expect very low class membership probability for any of the 8 classes. We have planned to do an outlier study and develop an anomaly detection model to identify these new sources.

Our next goal would be generate class membership probabilities table for all the point sources in the Chandra Source Catalog-2.0. We would publish this probabilistic classification alongwith the compiled multi wavelength data to a general public accessible portal. The portal will allow the user

to select the cone search region, a probability threshold, class or list of classes of their choice. The output will be all the sources lying in the selected region, which belong to the selected class with class membership probability more than the selected threshold. This portal will be of immense use for research in X-ray astrophysics.