RNAAS RESEARCH NOTES OF THE AAS

Machine Learning-based Classification of Variable Stars

To cite this article: Anjie Liu and Jasmine C. Xu 2024 *Res. Notes AAS* **8** 83

Research Notes of the AAS

Draft version March 13, 2024
Typeset using LaTeX default style in AASTeX631

Machine Learning-Based Classification of Variable Stars

Anjie Liu[1] and Jasmine C. Xu[1]

[1] *The University of Texas at Austin*

ABSTRACT

The classification of variable stars, essential for revealing information on stellar properties and cosmic distances, traditionally relied on statistical methods and limited data. With the emergence of transformative machine learning methodologies and large stellar surveys, we are able to perform more efficient, accurate, and robust handling of expansive databases. We deploy two different machine learning models - random forest and XGBoost - that effectively classify variable stars identified in the most recent phase of the Optical Gravitational Lensing Experiment (OGLE-IV) using four features: 1) time of minimum brightness, 2) I-band amplitude, 3) mean or maximum I-band magnitude, and 4) period. The two models achieve a cross-correlation score of 99.00%, indicating largely shared classifications. The results produced by the models uncover new insights into the precision of supervised learning models in predicting variable stars with the aid of the most up-to-date survey data.

Keywords: Variable stars — Machine Learning — Classification — Random Forest — XGBoost

1. INTRODUCTION

Since their discovery in 1638, variable stars have intrigued amateur and professional astronomers alike (Hogg 1933). Much is now known about variable stars—stars whose brightness fluctuates over time—including the fact that differences in factors such as period, luminosity, and mass split variable stars into several main classes and further subclasses. While human astronomers may be more capable of distinguishing noise from legitimate data, the feasibility of human classification of celestial objects is limited by the sharp increase in the pace and volume of astronomical data collection. Given this trend, a suitable approach to astronomical questions is to use modern technology that can handle vast amounts of data with reasonable efficiency and accuracy.

The primary goal of this work is achieved by building and comparing a random forest and an XGBoost model. Model performances can be quantified by computing and comparing classification metrics.

2. DATA ANALYSIS

To provide an accurate classification, we obtained data from the fourth and most recent phase of the Optical Gravitational Lensing Experiment (OGLE-IV) (Udalski et al. 2015). OGLE uses microlensing to identify stars in the Small and Large Magellanic Clouds, as well as in the Galactic Bulge and Galactic Disk, providing one of the most comprehensive and up-to-date databases of classified variable stars (Udalski et al. 1992). Additionally, OGLE-IV includes an extensive collection of features for each star, including the average I-band and V-band magnitudes, I-band amplitudes, and photometric data. This wealth of features makes the database well-suited as a source of input data for supervised classifier models.

Ultimately, we extracted a dataset comprised of around 736,000 variable stars taken from all target fields of the OGLE-IV database. This encompasses eight types of variable stars, namely, eclipsing binaries, RR Lyrae variables, long-period variables, Delta Scuti variables, classical Cepheids, type II Cepheids, Heartbeat stars, and anomalous Cepheids.

3. METHODOLOGY

We chose to build a random forest classifier and an XGBoost classifier, both of which are learning methods who combine the predictions of multiple learners to obtain high prediction accuracy (Breiman 2001). Our selected features to use in the training of our models are 1) time of minimum brightness in days, 2) I-band amplitude, or main eclipse depth, 3) mean (for pulsating stars) or maximum (for eclipsing binaries) I-band magnitude, and 4) period in days.

As previously mentioned in section 2, there is a significant amount of class imbalance present within our dataset that could greatly hinder a classifier model's ability to correctly identify minority classes. We aim to address this class imbalance using Synthetic Minority Oversampling Technique (SMOTE), a type of data augmentation that generates synthetic samples for the minority classes using the existing minority class samples (Chawla et al. 2011). This method will ensure that there are equal numbers of stars in each class, and ultimately increase the model's ability to learn from the minority classes.

Following the preprocessing of our dataset, we constructed a random forest model and an XGBoost model. The construction, training, and testing of our machine learning models relied heavily on Python's Scikit-learn library. To optimize model performance, a randomized search was used to tune hyperparameters for both models. The determined optimal parameters were then used to train the respective models.

4. RESULTS

Our random forest and XGBoost classifiers achived accuracy scores of 97.71% and 97.68%, respectively. Their confusion matrices, shown in **Figure 1**, display comparable overall trends in that for both models, the main source of error comes from the misclassifications of RR Lyrae as eclipsing binaries and vice versa. By examining this pattern more closely, we observed that in many instances of these misclassifications, the features were extremely similar, such that the model could not distinguish between them. Since only 4 features were used to train our models, it might be beneficial for future investigations to use more features.

Figure 1. Confusion matrix of random forest classifier

To understand the concordance in classifications made by our models, we computed a cross-correlation score of 99.00%. This suggests that a vast majority of classifications overlap between both models. In the subset of classifications that differed, many of them involved misclassifications of RR Lyrae and eclipsing binaries.

5. REFLECTION AND FUTURE DIRECTIONS

Our random forest and XGBoost models achieve high accuracy scores of 97.71% and 97.68%, respectively. These satisfactory results across all variable star types are largely aided by the choices of features, as well as the effective management of imbalanced data.

While our investigation has shown positive results, we hope to extend it by constructing more machine learning models that can be compared to our current random forest and XGBoost classifiers. We most strongly aim to take advantage of the availability of photometric data to implement a recurrent neural network. Unlike our current models, which rely on summary features, neural networks can directly take in raw data, such as photometric readings, and so would provide a new approach to variable star classification. Similarly, a convolutional neural network could be applied alongside this, given that photometric data can be represented as images.

Another further step we aim to take is to apply our current models to other variable star databases in addition to OGLE-IV, which our models were trained upon. If applied to classified datasets, this would allow us to confirm the applicability and effectiveness of our models; and if applied to unclassified datasets, this gives us the potential of identifying new variable stars.

ACKNOWLEDGEMENTS

We would like to thank Dr. Shyamal Mitra, Dr. Karl Gebhardt, and our peer mentors, Jose Ordonez and Rik Ghosh, for their support and advice throughout our research journey. We would also like to extend our gratitude to the Geometry of Space stream of the Freshman Research Initiative at The University of Texas at Austin for giving us the opportunity and resources to explore our interests and make scientific contributions.

REFERENCES

Breiman, L. 2001, Machine Learning, 45, 5

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. 2011, doi:10.48550/ARXIV.1106.1813

Hogg, E. G. 1933, Journal of the Royal Astronomical Society of Canada, 27, 75, bibliographic Code: 1933JRASC..27...75H

Udalski, A., Szymanski, M., Kaluzny, J., Kubiak, M., & Mateo, M. 1992, Acta Astronomica, 42, 253, 1992AcA....42..253U

Udalski, A., Szymański, M. K., & Szymański, G. 2015, OGLE-IV: Fourth Phase of the Optical Gravitational Lensing Experiment, arXiv, arXiv:1504.05966 [astro-ph]