Creating an Automated Classifier to Approximate the By-eye Selection of HAT Exoplanet Candidates

Will Coulton
Department of Physics
Princeton University
wcoulton@princeton.edu

Joshua Wallace
Department of Astrophysical Sciences
Princeton University
joshuajw@princeton.edu

Abstract

A proposal to use machine learning techniques to aid in the binary classification of data from the Hungarian Automated Telescope (HAT) arrays. The HAT arrays are designed to detect periodic dimmings of stars, potentially caused by orbiting planets around those stars. There are many signals, both physical and spurious, that can approximate planet signals, so it is important to weed out as many false positives as possible. After an automated pipeline removes probable false planets, a final manual by-eye selection is made to create the final set of planet cadidates. We propose to train classifier models to approximate this by-eye selection with the hope of finding a sufficiently well-performing model that can stand as a proxy for or even entirely replace the manual portion of the selection pipeline.

1 Introduction

Since the discovery of the first extra-solar planet (exoplanet) in the early 90's [6], nearly 3000 exoplanets have been discovered [3], with potentially thousands more exoplanets remaining unconfirmed or as yet undiscovered in existent data. This explosion in exoplanet discovery has been fueled by the funding and construction of many exoplanet detection surveys. The majority of exoplanets have been found by the space-based *Kepler* telescope and its primary and "K2" surveys [2], [5]. Ground-based surveys, while not as sensitive to smaller planets as *Kepler* (due to atmospheric distortions and other problems associated with observing from the ground), are sufficiently sensitive to discover Jupiter-sized planets on short-period orbits. (Such planets are called "hot Jupiters" because they are the same size as Jupiter but are sufficiently close in to the stars they orbit (much closer than Mercury is to our sun) that the star heats them up to several thousands of degrees Celsius.) Since ground-based surveys are much cheaper to run per area of sky monitored than spaced-based surveys, large-scale ground-based surveys have discovered more hot Jupiters than space-based surveys.

Of the few hundred hot Jupiters that have been found to date, the plurality ($\sim$100) have been found by Princeton's own Hungarian Automated Telescope (HAT) collaboration, headed by Prof. Gaspar Bakos of the Department of Astrophysical Sciences. This collaboration runs two surveys, both of which use telescopes that are no bigger than large camera lenses. The longest-running of the two surveys is HAT-Net, a collection of five telescopes in Arizona and two telescopes in Hawaii [4]. The other survey is HAT-South, which consists three locations with eight telescopes each: Chile, Namibia, and Australia [1].

Planets are far too faint compared to their host stars to be able to detect directly. Instead, what the HAT arrays do is monitor the brightness of stars as a function of time. When a planet crosses in front the star it orbits, the amount of light detected from that star is decreased. When a star is found to have periodic dips in brightness, it is labelled as a potentially planet-hosting star. Various cuts are made to help ensure that the signal is neither spurious nor caused by another astrophysical source other than a planet. Most of these cuts are automated, but the final step in the HAT pipeline is a

manual, by-eye examination of the signal. Planet candidates that pass this step then are sent for follow up (and hopefully confirmation!) at bigger telescopes.

Other than this manual step, the vetting of possible planet candidates is completely automatic. This single manual step is quite labor intensive, and prevents a fully automatic characterization of planet detection efficiency, which is necessary for the calculation of statistics related to the occurrence rate of these planets. Thus, it would be nice to have an automated approximation to the by-eye work that has occurred. This is the purpose of this project.

2 Related Work

3 Data

The data consists of $\sim$1000 examples of potential planet candidates that were decided, by eye, to be included as planet candidates, and $\sim$30,000 examples of potential planet candidates that were decided to not be regarded as planet candidates. There are $\sim$90 features, all continuous, for each potential planet candidate. These features include parameters and statistics related to the fitting of the brightness dips, parameters related to the measured planet size and orbital period, and properties of the host star (how bright it is in different colors, etc.). The data were extracted from the proprietary HAT database using a script provided by Joel Hartman, Research Scienctist at Princeton University. The database itself has not reference currently.

4 Methods

Our first approach to this problem will be to investigate the simple binary classifiers that we have covered in class so far. These include Naive Baye's, random forests, support vector machine (SVM), and logistic regression. With logistic regression will investigate two types of regularization using l_1 and l_2 penalties. We will also explore a set of feature selection methods such as mutual information.

Whilst our main goal is to predict which objects would be manually selected, we are also interested in running a set of unsupervised learning algorithms to view any hidden structure in the data set. These methods will include K-means and gaussian mixture models. This will be useful for potentially identifying new approaches to automating this process as well as potentially probing some new relationships.

5 Results

6 Discussion and Conclusion

Acknowledgments

We are grateful to Joel Hartman for providing a script to extract the necessary data from the HAT database. We are also grateful to the entire HAT team (PI: Gaspar Bakos) for their many dedicated years of constant observations and the huge pile of astronomical data they've collected.

References

[1] G Bakos, C Afonso, T Henning, A Jordán, M Holman, RW Noyes, PD Sackett, D Sasselov, Gábor Kovács, Z Csubry, et al. Hat-south: a global network of southern hemisphere automated telescopes to detect transiting exoplanets. *Proceedings of the International Astronomical Union*, 4(S253):354–357, 2008.

[2] William J Borucki, David Koch, Gibor Basri, Natalie Batalha, Timothy Brown, Douglas Caldwell, John Caldwell, Jørgen Christensen-Dalsgaard, William D Cochran, Edna DeVore, et al. Kepler planet-detection mission: introduction and first results. *Science*, 327(5968):977–980, 2010.

[3] exoplanets.org. Exoplanets.org main page. http://exoplanets.org/. Accessed: 2017-04-21.

[4] JD Hartman, G Bakos, KZ Stanek, and RW Noyes. Hatnet variability survey in the high stellar density kepler field with millimagnitude image subtraction photometry. *The Astronomical Journal*, 128(4):1761, 2004.

[5] Steve B Howell, Charlie Sobeck, Michael Haas, Martin Still, Thomas Barclay, Fergal Mullally, John Troeltzsch, Suzanne Aigrain, Stephen T Bryson, Doug Caldwell, et al. The k2 mission: Characterization and early results. *Publications of the Astronomical Society of the Pacific*, 126(938):398, 2014.

[6] A. Wolszczan and D. A. Frail. A planetary system around the millisecond pulsar PSR1257 + 12. *Nature*, 355:145–147, January 1992.