Automatic Detection of Microlensing Events in the Galactic Bulge using Machine Learning Techniques

Selina Chu,[1], Kiri L. Wagstaff[1], Geoffrey Bryden[1] and Yossi Shvartzvald[2]
[1]*Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA;* `Selina.Chu@jpl.nasa.gov`

[2]*IPAC, Science & Data Center for Astrophysics & Planetary Sciences, California Institute of Technology, Pasadena, CA, USA*

Abstract.
The Wide Field Infrared Survey Telescope (WFIRST) is a NASA flagship mission scheduled to launch in mid-2020, with more than one year of its lifetime dedicated to microlensing survey. The aim is to discover thousands of exoplanets via their microlensing lightcurves, which will enable a Kepler-like statistical analysis of planets 1-10 AU from their parent stars and revolutionize theories of planet formation. The goal of our work is to create an automated system that has the ability to efficiently process and classify large-scale astronomical datasets that missions such as WFIRST will produce. In this paper, we discuss our framework that utilizes feature selection and parameter optimization for classification models to automatically differentiate the different types of stellar variability and detect microlensing events. The use of feature selection enables us to learn which characteristics distinguish the different types of events and to classify high-dimensional data more efficiently. We demonstrate our proposed method on datasets acquired from UKIRT's wide-field near-IR camera that surveys the galactic bulge.

1. Introduction

The Wide Field InfraRed Survey Telescope (WFIRST) flagship mission Spergel et al. (2015) is scheduled to launch in mid-2020's with > 1 year of its lifetime dedicated to a microlensing survey. This survey will discover thousands of exoplanets near or beyond the snowline via their microlensing light curve signatures, enabling a Kepler-like statistical analysis of planets at 1-10 AU from their host stars and potentially revolutionizing our understanding of planet formation.

In preparation for WFIRST's microlensing survey, NASA's Exoplanet Program Analysis Group *ExoPAG* chartered a Study Analysis Group $SAG-11$ identified three programs for the microlensing community to undertake to prepare for WFIRST Yee et al. (2014). Our work aims to directly address key precursor requirements, which includes identifying target fields and developing analysis tools for WFIRST.

Using machine learning to modernize microlensing event detection is the critical next step both for the UKIRT survey now and for the success of the WFIRST flagship mission in the future. Machine learning builds a predictive model from human-labeled examples that can 1) detect events of interest, such as microlensing, and 2) characterize the properties of these events.

2. Microlensing

For any microlensing survey, the number of detected planets is proportional to the number of microlensing events, which in turn depends on the density of observable stars. Microlensing surveys have therefore concentrated their observing efforts toward the Galactic bulge, where the stellar surface density is highest. However, these surveys have been traditionally conducted at optical wavelengths, which suffer from high extinction from dust near the Galactic plane and center. To maximize the microlensing event rate, their fields were selected based on a balance between stellar surface density and dust extinction. Observing in the near-infrared (NIR) mitigates the effects of high extinction, enabling observations closer to the Galactic center, where the event rate is expected to be higher. Understanding this potential, WFIRST will conduct its microlensing survey in the NIR. However, until re- cently, no dedicated NIR microlensing surveys have been conducted, and so the event rate in the NIR âĂŞ which is crucial for WFIRST field optimization âĂŞ has not been measured. For this reason, ExoPAG SAG-11 identified a NIR microlensing survey as a key precursor activity for WFIRST (Yee et al. (2014).

3. Data

In order to map the unknown NIR event rate, NASA is funding a NIR survey with the United Kingdom Infrared Telescope (UKIRT), a 3.8-m telescope on Mauna Kea in Hawaii. Initially started as a pilot study in support of the 2015 Spitzer microlensing campaign, its target fields were relocated in 2016 to match Kepler's K2C9 dedicated microlensing campaign. In 2017, the program was expanded and the fields changed again, now to cover all potential WFIRST fields, including the Galactic center, which is inaccessible to optical surveys due to the high extinction

From our initial analysis of this data we have successfully detected the first five microlensing events ever discovered in the NIR Shvartzvald et al. (2017) and found a Jupiter-mass planet orbiting at 4 AU around a solar-type star, the first microlensing exoplanet discovered and characterized independently of the long-standing optical surveys Shvartzvald et al. (2018). While we have subsequently identified many additional unpublished events, our analysis is based on a by eye selection of the most significant events. Many additional events lie below our conservative detection threshold. More importantly, the detection efficiency has not been calculated, without which it is impossible to measure and map the underlying rate of microlensing event. For this project we will develop statistically rigorous, computationally efficient tools to allow us to extract the full potential from the UKIRT dataset.

3.1. Feature Extraction

To identify each lightcurve, we derived features using a grid-based approach for microlensing fit, based on method proposed in **?**. The model grid uses two parameters: effective event timescale and time when event peaks. To derive analytically the best fit, each model lightcurve is scaled by the source flux and blended flux. We compare each microlensing fit against a straight line to find the goodness of fit. The goodness of fits provides a simple way to identify lightcurves with significant variability and are used to perform intial selection of microlensing candidates. From this process, 66 features are derived from statistics evaluating each lightcurve. Examples include statistics from microlensing fit and event timescale, event impact parameter, source flux, blend flux, etc

4. Classification Methods

We are building classifier based on the extracted features with focus on identifying microlensing events more effectively. To evaluate the performance of our recognition system, we examined the following three classification methods: Random Forest, and Support Vector Machine (SVM), K- Nearest Neighbor (KNN). In addition, we also used model selection to optimize model parameters and feature selection to improve class discrimination.

The first step in machine-based classification is to supply the computer with a set of true classifications to train on. So while our eventual goal is to have the computer carry the load for the overall dataset, we must still perform many (thousands) of individual by-eye classifications. To ease this process, we have developed the Manual UKIRT Lightcurve Eval- uator (MULE), a python-based GUI application that allows users to quickly cycle through and evaluate a set of stored light curve

4.1. Feature Selection

One of the problems in using a large number of features is that there are many potentially irrelevant features that could negatively impact the quality of classification. In using feature selection techniques, we can choose a smaller feature set to reduce the computational cost and running time, as well as achieve an acceptable, if not higher, recognition rate. Adding more features is not always helpful; as the feature dimension increases, data points become more sparse and some features are essentially noise. This leads to the issue of selecting an optimal subset of features from a larger set of possible features that will yield the most effective subset. The optimal solution is using an exhaustive search of all the features. This requires $2^{66} - 1$, or roughly 10^19 combinations. Instead of performing 10^19 computations, we use a greedy search for selecting the features. There are various ways of performing feature selection, such as forward feature selection, backward selection, branch and bound, and stochastic search, each with its advantages and disadvantages. We used forward feature selection for our experiments since it is simple and straightforward

5. Experimental Setup and Results

Model selection is performed at each iteration of the feature selection process to learn the optimal parameters for each classifier type and feature set. We used the filtered dataset from 2017 UKIRT survey, we train a 3-class classifier to differentiate between Microlesning, Variable, and Glitch. We used 3-fold cross validation, spliting the data into 2/3 for training and 1/3 for test, over three trials. Data are normalized using zero mean and unit variance

SVM and Random forest results are similar with RF performing slightly better overall. We focus on using Random forest, as it is more robust, with the following selected model parameters and 49 number of features.

Tested the classifier with the learned model θ on the unlabeled test candidate lightcurves Figure (right) shows some examples of lightcurves found from the unlabeled test candidates as microlensing (m) event, with class probability of being in the microlensing class, $p(m) \geq 0.8$. Each lightcurve plot is ranked from highest to lowest (left to right, top to bottom)

Figures **??**

6. Conclusions and Future Work

This paper investigates techniques for developing a microlensing detection system using using the proposed features. The classification system was successful in classifying the different types of events. We also found that using high number of features is not always beneficial to classification. In using forward feature selection, a form of greedy search, a smaller subset were required to achieve a high recognition rate. We have also identified features that has potential to detect microlensing event with high probability. Currently we are developing a framework for event injections to evaluate detection efficiency. We are also using active learning to reduce manual labeling of lightcurves by automatically selecting most informative unlabeled lightcurves to improve classification performance. In parallel, we are so investigating into domain adaptation to transfer findings from UKIRT to WFIRST more effectively

Currently we are developing a framework for event injections to evaluate detection efficiency. In parallel, we are also developing a tool using active learning to reduce manual labeling of lightcurves by automatically selecting most informative unlabeled lightcurves to improve classification performance. The next step, we are planning to learn a model domain adaptation to transfer findings from UKIRT to WFIRST more effectively.

7. References

References

Shvartzvald, Y., Bryden, G., Gould, A., Henderson, C. B., Howell, S. B., & Beichman, C. 2017, Astron. J., 153. URL `http://dx.doi.org/10.3847/1538-3881/153/2/61`

Shvartzvald, Y., Novati, S. C., Gaudi, B. S., Bryden, G., Nataf, D. M., Penny, M. T., Beichman, C., Henderson, C. B., Jacklin, S., Schlafly, E. F., & Huston, M. J. 2018, Astrophys. J. Lett., 857. `arXiv:1802.06795v2`

Spergel, D., Gehrels, N., Baltay, C., Bennett, D., Breckinridge, J., Donahue, M., Dressler, A., Gaudi, B. S., Greene, T., Guyon, O., Hirata, C., Kalirai, J., Kasdin, N. J., Macintosh, B., Moos, W., Perlmutter, S., Postman, M., Rauscher, B., Rhodes, J., Wang, Y., Weinberg, D., Benford, D., Hudson, M., Jeong, W.-S., Mellier, Y., Traub, W., Yamada, T., Capak, P., Colbert, J., Masters, D., Penny, M., Savransky, D., Stern, D., Zimmerman, N., Barry, R., Bartusek, L., Carpenter, K., Cheng, E., Content, D., Dekens, F., Demers, R., Grady, K., Jackson, C., Kuan, G., Kruk, J., Melton, M., Nemati, B., Parvin, B., Poberezhskiy, I., Peddie, C., Ruffa, J., Wallace, J. K., Whipple, A., Wollack, E., & Zhao, F. 2015, ArXiv e-prints. `1503.03757`

Yee, J. C., Albrow, M., Barry, R. K., Bryden, G., Chung, S.-j., Gaudi, B. S., Gehrels, N., Gould, A., Penny, M. T., Rattenbury, N., Skowron, J., Street, R., & Sumi, T. 2014, 1. `arXiv:1409.2759v1`