Using **statistical analysis** as a way of pre-classifying data **significantly reduces the amount of manual classification** and **provides meaningful information about the data.**

Leveraging Statistical Analysis to Develop Classification Labels for Time Series Data

E. L. Howard[1], K. R. Covey[1], J. R. A. Davenport[2]

1: Western Washington University, 2: University of Washington

Introduction

Manually classifying thousands of data plots is exhausting. Machine learning (ML) can take on this burden, but a robust model requires classification labels to train on, which brings us back to the same problem. We propose a programmatic pre-classification using statistical analysis.

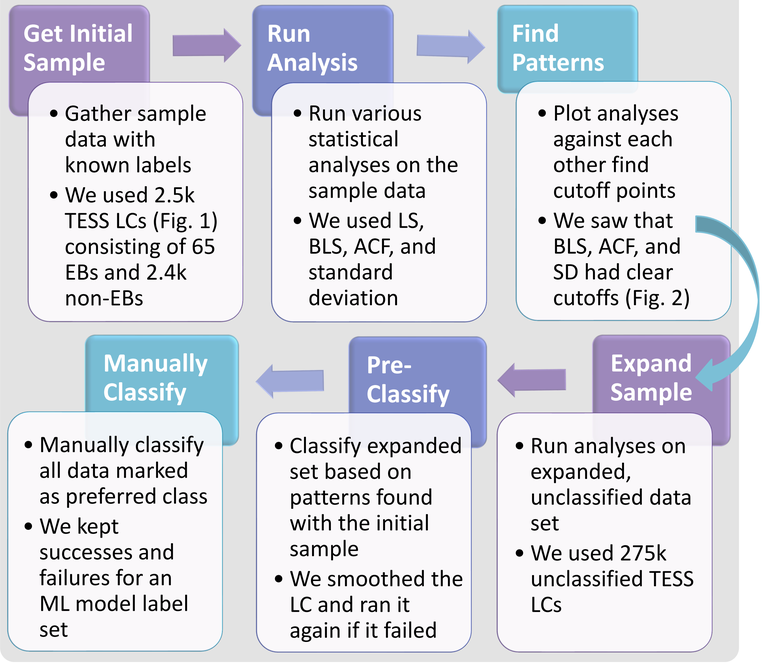
Using statistical analysis, we can find patterns in our data that allow us to pre-classify hundreds of thousands of plots in a matter of hours.

Types of Statistical Analyses

- **Lomb-Scargle (LS)** using `AstroPy` compares the data to a sinusoidal wave.
- **Autocorrelation Function (ACF)** using `exoplanet` compares the data to itself.
- **Box-Least Squares (BLS)** using `AstroPy` compares the data to a square wave.

All analyses use power. Power correlates with goodness of fit. Higher the power, better the fit.

Methods



Sample Application

We chose light curves from the Transiting Exoplanet Survey Satellite (TESS), looking for Eclipsing Binaries (EBs). EBs make up roughly 2% of the entire data set—which makes manual classification difficult.

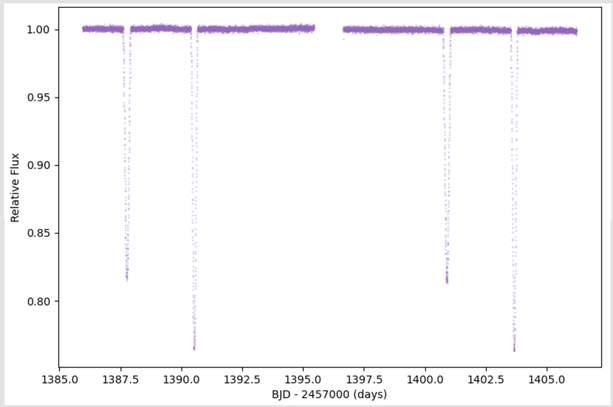


Figure 1: An example of a TESS[1] light curve (LC) showing an EB.

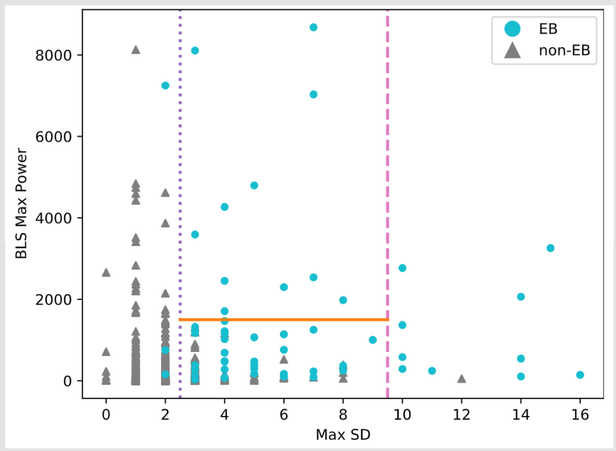


Figure 2: A sample of how the statistical patterns were determined with BLS Max Power vs. Max SD. Anything above the solid orange line and to the right of the dashed pink line – predominantly EBs – were labeled as EBs. Anything to the left of the dotted purple line – predominantly non-EBs – were labeled non-EBs. Anything else was smoothed and ran through the classifier again.

Sample Results

- We ran pre-classification on 275k LCs
  - Took 5 hours to complete
  - Found 500 objects with at least one EB flag consisting of 3.1k LCs to manually classify
    - 2k properly classified EBs
    - 80 false positives
    - 500 properly classified non-EBs
    - 540 false negatives
- The complete label set to be used in a machine learning model, after augmentation, consists of:
  - 5k EBs
  - 7k non-EBs

Future Work

This research will be reapplied to TESS light curves in order to classify extrasolar flares.

Acknowledgments

This work was supported in part by the Distributed Research Experiences for Undergraduates (DREU) program, a joint project of the CRA Committee on the Status of Women in Computing Research (CRA-W) and the Coalition to Diversify Computing (CDC), which is funded in part by the NSF Broadening Participation in Computing program (NSF592BPC-A #1246649).

1: This project used data collected with the TESS mission, obtained from the MAST data archive at the Space Telescope Science Institute (STScI). Funding for the TESS mission is provided by the NASA Explorer Program. STScI is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract NAS 5–26555.

CRA-WP
Computing Research Association
Widening Participation