Draft version December 7, 2020
Typeset using LATEX **twocolumn** style in AASTeX63

Leveraging Statistical Analysis to Develop Labels for Astronomical Time Series Data

Erin Howard [1] and James R. A. Davenport [2]

[1] *Western Washington University*
[2] *University of Washington*

ABSTRACT

Developing an effective machine learning training set for classifying new time domain astronomy data sets can be a daunting task, particularly for a classification that has many representations. Utilizing statistical analysis to pre-classify a portion of the data can lessen the burden of manually labeling data. After developing a model on a sample set of $\sim 2,500$ with known binary classification labels, it was applied to a data set of $\sim 275,000$ with unknown classification. This resulted in $\sim 2,000$ light curves classified as eclipsing binaries (EBs) consisting of 500 objects. All light curves belonging to the 500 objects were manually analyzed and $\sim 2,400$ were properly classified as true EBs and ~ 520 were classified as non-EBs. Adding in the preliminary sample set of manually classified LCs resulted in a training set of $\sim 2,400$ true classifications and $\sim 3,000$ false classifications that can be used to train a machine learning model that would classify the entire data set of $\sim 275,000$ LCs.

1. INTRODUCTION

A machine learning algorithm that classifies data is only as good as its training set, but developing a set of classification labels from an expansive data set manually can be daunting. This is particularly true when potentially less than 2% of the total data set contains the desired classification, as is the case with eclipsing binary (EB) stars.

From May of 2009 to May of 2013, the Kepler mission observed $\sim 200,000$ stars and identified ~ 3000 EBs. (Abdul-Masih et al. 2016) If this number is representative of the entire population of stars that we can see from Earth, then fewer than 2% of the stars observed will end up being EBs. Kepler's successor, the Transiting Exoplanet Survey Satellite (TESS), will have 350 times more sky coverage than Kepler over the course of its mission which has already resulted in an expansive data set that does not yet have a catalog of EBs.

Filtering through a large data set by hand to build a classification training set, where only a few percent of the total objects are of the desired classification, is tedious and time intensive. While others have overcome this by simulating data, it also requires a large number of examples or else the generated samples will be too close to the originals to truly capture the variety in the expanded data set. Rather than generating fake data, a portion of the original data can be manually classified and then analyzed for patterns using statistics. Once the patterns are found, a statistical model can be run on a more expansive portion of the data set in order to provide classification labels. This method not only provides a larger training set faster than manually vetting, but the patterns that are found when building the classifier can also provide avenues for further research to explore.

Using light curves (LCs) from the TESS data set as a representative of univariate time series data, we found patterns in $\sim 2,500$ manually classified LCs using autocorrelation and box least squares analyses. We developed this into a binary classifier that allowed us to determine ~ 500 objects that contained at least one classified EB, which were comprised of $\sim 3,100$ LCs in total. This is a reasonable amount of data to manually classify, and a significant portion of them are of the desired class. This ensures that the training set is not dominated by the majority class like it would without the statistical classifier boosting EB representation.

Through manual classification we identified $\sim 2,000$ properly classified EBs, ~ 50 Type I errors and relabeled them as non-EBs, ~ 500 properly classified non-EBs, ~ 500 Type II errors and relabeled them as EBs, ~ 30 ambiguous LCs that were classified as EB, and ~ 40 ambiguous LCs that were classified as non-EB. Ambiguous LCs were not included in the training set. It is important to note that even if the object is an EB, if the particular section did not contain an eclipse, it was labeled as a non-EB.

Corresponding author: Erin Howard
howarde7@wwu.edu

After manual classification, the training set consisted of $\sim 2,500$ EBs and ~ 550 non-EBs. After incorporating the original sample data, our training set consisted of $\sim 2,500$ EBs and $\sim 3,500$ non-EBs. We increased the training set through data augmentation by reversing the values which doubled our training set for a final training set size of $\sim 5,000$ EBs and $\sim 7,000$ non-EBs.

2. METHODOLOGY

2.1. *TESS Light Curves*

2.1.1. *Initial Data Samples*

The initial data samples that were used to find statistical patterns were originally selected to include known EBs. The first set contained 11 EBs and 55 non-EBs, the second set contained 2 EBs and 254 non-EBs. The third set was gathered at random from sector 25 and contained 12 EBs and 1023 non-EBs. The fourth set was gathered at random from sectors 1 through 13 and contained 39 EBs and 1106 non-EBs.

The final data sample that was used to develop the classifier consisted of 64 EBs and $2,438$ non-EBs from 16 of 26 sectors.

2.1.2. *Data Selection*

The data we used was the Pre-search Data Conditioning SAP (PDCSAP) flux and time taken from the LC table inside the Fitz file for the object's sector data. We eliminated all data points that were flagged by TESS to have quality issues and clipped ~ 1.5 days on either side of data gaps, including the start and end of the sector's observation, to eliminate systemic errors caused by the satellite moving into position. After the data was cleaned, we divided the flux by the median flux and looked at the resulting relative flux values for the classifier.

2.1.3. *Smoothing and Rolling Window*

In order to flatten the dominant curves in LCs that are caused by natural variability and/or starspots, we performed two smoothing operations, with an additional classification attempt in between, to the LCs that were left unclassified after the initial classification attempt. The smoothing was done by a rolling median using a data-dependent window.

The window used for the rolling smooth was determined by a combination of the number of peaks reported by the ACF function divided by the median difference in time between two consecutive flux measurements. This value was then divided by a window factor. The window scalar was selected by testing a range of scalars from 2 to 10 and comparing their success with correctly classifying the sample data. The window factor that was chosen identified the most EBs and misclassified the least amount of non-EBs.

After the rolling median was calculated, it was then subtracted from the original data that smoothed the short term fluctuations in data. This helped the ACF and BLS functions determine periods with higher statistical power ratings that were better able to correctly identify EBs.

2.2. *Statistical Analysis*

2.2.1. *Lomb-Scargle*

The Lomb-Scargle (LS) analysis compares the data to a sinusoidal wave. The more the sinusoidal wave fits the data, the higher the power. The less the sinusoidal wave fits the data, the lower the power. For this analysis we used the LS analysis from the Astropy package in Python using autopower to determine the frequencies in which to check based on the data. Once the frequencies and powers were generated, we found the frequency with the highest power and converted it to a period in days.

2.2.2. *Autocorrelation*

The autocorrelation function (ACF) takes a section of the data and moves it progressively through the data. The more the selected section matches the data, the higher the power. The less the selected section matches, the lower the power. For this analysis we used the ACF analysis from the Exoplanet package in Python and restricted ourselves to a minimum period of 0.05 days (72 minutes) and a maximum period of 27 days. This restriction was decided based on the minimum and maximum periods that the TESS observations can capture per sector.

2.2.3. *Box-Least Squares*

The box-least squares (BLS) analysis compares the data to a box wave. The more the box wave fits the data, the higher the power. The less the box wave fits the data, the lower the power. For this analysis we used the Astropy package in Python and restricted ourselves to three potential durations (20 minutes, 40 minutes, 80 minutes, and 144 minutes) and three periods based on the highest powered period from the ACF analysis (the original period, the half period, and twice the period). The BLS analysis in the Astropy package was the most complicated analysis and we made these duration and period choices in order to reduce computation time.

2.2.4. *Z-Score and Average Max*

In statistics, the z-score of a data point represents the number of standard deviations it is away from the mean. For the purposes of this classifier, we chose to find the z-score of all data points relative to the median relative flux. Since eclipses block the light coming from the binary stars, the resulting LC has a significant dip in relative flux where the eclipse takes place. Since we were only looking for decreases in flux, we compared the negative z-scores and isolated the minimum.

2.2.5. *Finding Cutoffs*

Lomb-Scargle, autocorrelation, and BLS power scores were plotted against each other as well as average max z-score. These plots were analyzed at all levels of smoothing in order to find distinct boundaries between EBs and non-EBs.

The Lomb-Scargle function did not have a clear division between EBs and non-EBs in any of the plots, nor did it have a maximum or minimum power value that divided EBs from non-EBs. This is likely because of the algorithm that determines the power of the Lomb-Scargle analysis. While some EBs have a sinusoidal pattern, not all of them do. The same can be said for non-EBs. Due to this, we removed it from our classifier in order to make computation time faster.

The ACF function had a clear division between EBs and non-EBs when plotted average max z-score. The BLS function had a clear division between EBs and non-EBs when plotted against average max z-score that separated nearly all EBs in the sample set from non-EBs. The BLS max power and average max z-score had a clear minimum and maximum values, independent of all other statistical values, that separated nearly all EBs in the sample set from non-EBs.

2.2.6. *Reducing Type I Errors and Finding Type II Errors*

In order to decrease the number of LCs that were incorrectly labeled as EBs, $\sim$ 1.5 day's worth of data points was removed from the data on either side of observational gaps of at least one day. This was due to the systematic errors in observations that typically presented at the start or end of an observation when the satellite was moving into position.

In order to locate a portion of the LCs that were incorrectly labeled as non-EBs, all LCs from all objects that had at least one classified EB were included in the second round of classification. Due to the periodic nature of eclipses, it is likely that Type II errors are more likely in LCs of objects where at least one sector was classified as an EB.

3. THE STATISTICS-BASED CLASSIFICATION MODEL

3.1. *Autocorrelation*

With the sample data set in all stages of smoothing, it was found that LCs with an ACF max power of less than 0.05 and an average max z-score of less than 4 were typically non-EBs.

3.2. *Box-Least Squares*

With the sample data set, pre-smoothing, it was found that all LCs that had a BLS max power of 1500 or greater was an EB. It was also found that all LCs in the sample data that had an average max z-score less than 7 and BLS max power of less than 100 were non-EBs. After the first round of smoothing, that cutoff was lowered to 850. After the second round of smoothing, that cutoff was lowered to 200. In all stages of smoothing, very few EBs had a BLS max power of less than 60.

3.3. *Average Max Z-Score*

With the sample data set, pre-smoothing, it was found that all light curves that had an average max z-score of 10 or greater was an EB. After the first round of smoothing, that cutoff was lowered to 8. After the second round of smoothing, that cutoff was lowered to 9.

4. CONCLUSION

Through statistical analysis, we were able to narrow down the 275,000 LCs to a reasonable 3,100 that could be manually classified. The resulting manual classification found a 1% EB rate, which is roughly half the expected EBs in the sample. It is likely that a machine learning model will be able to successfully find what the statistical model could not.

While not perfectly balanced, the final label set has roughly 40% EBs, which is far greater than the 2% expected in the data set.In addition to being balanced, the set also includes over 100 examples of non-EBs that are close enough in shape that they resulted in a Type I error during classification. A balanced label set with examples of false positives provides a thorough sample for a machine learning model to learn from.

REFERENCES

Abdul-Masih, M., Prša, A., Conroy, K., et al. 2016, The Astronomical Journal, 151, 101, doi: 10.3847/0004-6256/151/4/101