ASTRONOMICAL IMPLICATIONS OF MACHINE LEARNING

ARUN DEBRAY
RAYMOND WU

DECEMBER 13, 2013

Abstract. In this project we use supervised learning to develop a classifier for stellar lightcurves to detect whether they demonstrate the existence of exosolar planets. We use various features selection methods; in particular we will be using dynamic time warping to measure the similarity between two temporal sequences.

1. Introduction

The recent discoveries of planets around stars other than our own is among the most significant trends in astronomy today. Long debated by philosophers and physicists alike, no such planets were known until 1992, when two planets were discovered around a star called PSR B1257+12. In the two decades since then, over a thousand such planets have been discovered, diverse in many ways. Thanks to these discoveries, astronomers are learning more about planetary systems other than our own, responding to these questions about other solar systems and even how probable Earth-like life could be in the universe.

In this paper, we will use the following standard terminology.

- An *exosolar planet* is defined as in [1] to be a planet that orbits a star other than the Sun.[1] The standard definition of a planet has two kinds of ambiguity: very low-mass objects in our solar system, such as Pluto, were defined to be "dwarf planets," and the boundaries of this definition aren't entirely clear. Very high-mass planets, however, resemble very small stars; though they don't undergo hydrogen fusion, they look very much like a dim type of star called a brown dwarf. The boundary is somewhat arbitrarily delineated at 13 Jupiter masses. However, neither of these is a great concern in this paper: science is yet unable to detect Pluto-sized worlds around another star, so the low-mass ambiguity does not arise in this data, and the high-mass boundary is not as important: a classifier that discovers planets and small brown dwarfs is still useful. However, it will be helpful to distinguish these systems from eclipsing binaries (see below).
- *Planetary transit* is a method of exoplanet detection, In general, because planets are very dim relative to their bright host stars, they cannot be directly imaged, in the same way that it is difficult to detect a firefly near a searchlight from afar. Thus, several indirect methods exist. Planetary transit repeatedly checks the brightness of a star over time; periodic, regular dips in this output sometimes happen because an exoplanet crosses between its sun and the observer. Thus, a planet may be detected without direct observation. See [2] for more information on transiting exoplanets.
- A *lightcurve* is a graph of a star's brightness over time. A transiting exoplanet will thus manifest itself as a lightcurve that is relatively constant, but with regular, small dips corresponding to the transits. See Figure 1 for two examples. The brightness is often given in units of magnitude rather than strict luminosity, because the logarithmic magnitude scale is generally easier to work with.
- An *eclipsing binary* is a pair of stars that orbit each other, but such that each eclipses the other from the Earth's point of view during the orbit. These generally don't contain transiting exoplanets, but instead form an important negative example. Their lightcurves look like those of exoplanets, but they aren't exoplanets.

Until recently, most exoplanets weren't detected by transit; astronomers used any of several other methods to find them. However, when the Kepler telescope was launched, it provided a wealth of data about transiting exoplanets, in particular showing that many stars could be surveyed at once. Since Kepler provided such a wealth of data about exoplanets, we decided to try to train a classifier on its lightcurves.

2. Methodology

We obtained Kepler light curves from the Mikulski Archive for Space Telescopes (see [4]). Lightcurves are stored in the `.fits` file format, so we used the AstroPy Python library, found at [6], to parse them. We first divided light curves into those corresponding to exoplanets and those corresponding to non-exoplanets. The non-exoplanets contained "non-variable" stars, which had relatively uniform light curves as well as stars with variable light curves

[1] The words *exoplanet*, *exosolar planet*, and *extrasolar planet* all mean the same thing, and are used interchangeably.

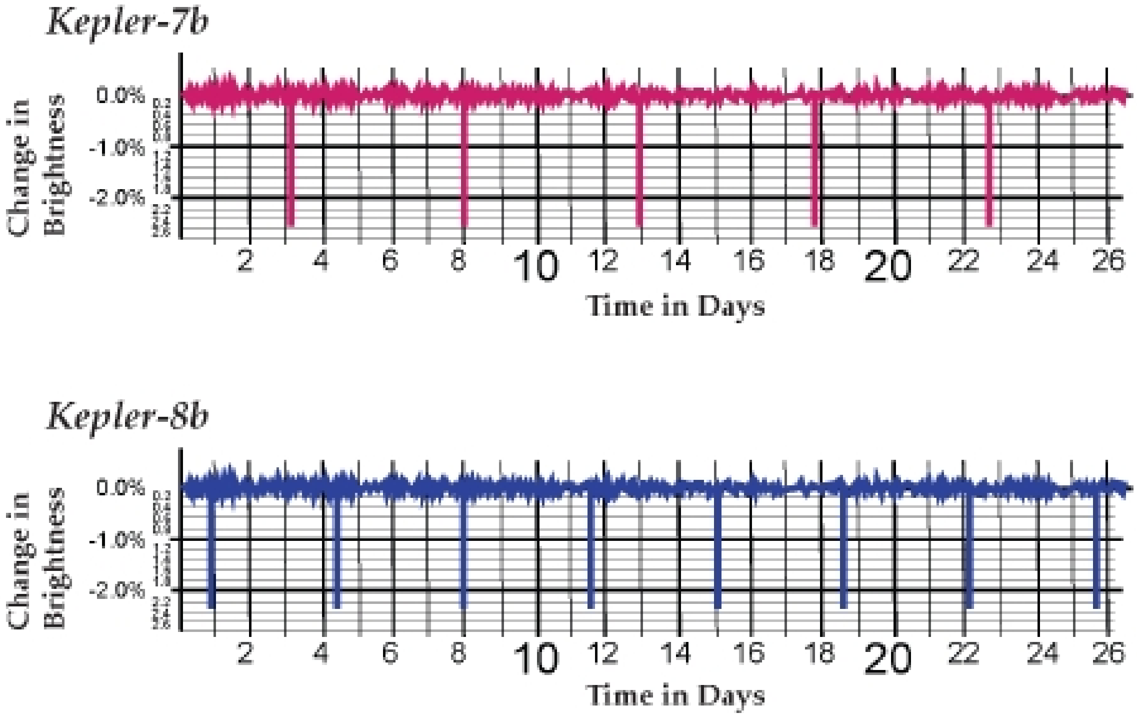


Figure 1. Lightcurves (graphs of stellar brightness versus time) of two planetary transit systems discovered by Kepler. The regular small dips correspond to the planet passing in front of its sun and slightly diminishing the observed magnitude. Source: [8]

that did not have exoplanets. These variable light curves were caused by other reasons, including eclipsing binary systems as well as intrinsically variable stars. The light curves are represented as time series data, which means that each light curve is represented by a series of attributes at each time point within some range. The attributes of the light curve stored in the binary table include:

- `TIME`: The time at the midpoint of the light curve
- `SAP_FLUX`: Simple aperture photometry flux in units of electrons per second contained in the optimal aperture pixels
- `SAP_FLUX_ERR`: The error in SAP flux in electrons per second
- `SAP_BKG`: The total background flux summed over the optimal aperture
- `SAP_BKG_ERR`: The error in the background flux
- `PDCSAP_FLUX`: Flux after Presearch Data Conditioning (PDC) has accounted for systematic error sources such as drift or focus change
- `PDCSAP_FLUX_ERR`: The error in PDCSAP flux
- `PSF_CENTR1`: The column centroid obtained by fitting the point spread function
- `PSF_CENTR1_ERR`: The error in PSF centroid 1
- `PSF_CENTR2`: The row centroid obtained by fitting the point spread function
- `PSF_CENTR2_ERR`: The error in PSF centroid 2
- `MOM_CENTR1`: The column value for the flux weighted centroid (first moment)
- `MOM_CENTR1_ERR`: The error in MOM centroid 1
- `MOM_CENTR2`: The row value for the flux weighted centroid (first moment)
- `MOM_CENTR2_ERR`: The error in MOM centroid 2
- `POS_CORR1`: The column position correction based on bright stars
- `POS_CORR2`: The row position correction based on bright stars

We decided early on that we should build a classification model on features extracted only from a couple of these attributes since not all the attributes were relevant to our query. We decided to go with attributes `SAP_FLUX`, `SAP_BKG`, `PDCSAP_FLUX`, `MOM_CENTR1`, `MOM_CENTR2`, `POS_CORR1`, and `POS_CORR2`. These values in particular were always defined over the entire time interval and provided important data about the light curve.

Since our lightcurves are time series, which are large series of data points separated by a uniform time interval, it is both extremely difficult and unhelpful to train any sort of classifier over the raw data points. Instead, we will need to extract features from each time series. These features have to be both manageable and descriptive of the light curve.

We decided to try a variety of feature selection methods. Initially, we decided to look for global features such as the arithmetic average, maximum, and minimum values for each attribute. These global features provide very high-level, general descriptions of the time series. Thus they provide a good set of initial features, but can only provide a limited amount of detail. Using metafeatures can be used to provide a more accurate description of the time series, but we decided to opt for a different type of feature.

Given that we were examining light curves, which vary in time and speed due to differences in the size and shape of the orbiting planets as well, differences in orbital periods, and differences in the instrument recordings, we want to be able to detect common patterns that are indicative of exoplanets regardless of these differences between different planets and stars. So to do this, we turned to an algorithm called dynamic time warping. Dynamic time warping (DTW), outlined in [5], calculates the optimal match between two time series by calculating a measure of similarity independent of the time, speed, and acceleration of the time series.

We decided to use dynamic time warping to calculate the similarity of each light curve with a "baseline" light curve. To find an appropriate baseline light curve, we took a relatively flat light curve from our training set of light curves without exoplanets. Then we used dynamic time warping to calculate a measure of similarity between each light curve and this baseline. We used this result as one of our features in addition to our global features.

After extracting all of our appropriate features, we needed to train a classifier. We used an open source machine learning algorithm suite called Weka ([3]), developed at the University of Waikato in New Zealand. We used a variety of classification methods including logistic regression, Naïve Bayes, support vector regression (SVR), and multilayer perceptron. We found that SVM did not work particularly well, so we used a modified SVM algorithm called RegSMOImproved[7]. In particular, we used a polynomial kernel to map the data to a higher dimension and then use the altered SMO algorithm to calculate a regression model. We also made use of the multilayer perceptron, an artificial neural network model that learns through backpropgation.

We tested our classifiers on our data sets using 10-fold cross validation. For testing the features generated by DTW, we had 150 light curves that were either confirmed exoplanets or were labeled as exoplanets but unconfirmed as of August 2013, and 103 light curves that were confirmed to not be exoplanets. So in total our training set size was 253 instances. For the means-of-attributes features, we had about 200 data points, split evenly between those with confirmed planetary transits and no known transits.

3. Results

See Table 1 for the overview of results; in general, a variety of algorithms performed quite well on this data, averaging an about 80% classification accuracy. We can take a look more specifically at how our logistic regression

Model	Correct	Incorrect	Mean absolute error	RMS error	Confusion matrix
Logistic (means of attributes)	159 (79.5%)	41 (20.5%)	0.2548	0.3341	$\begin{bmatrix} 90 & 10 \\ 31 & 69 \end{bmatrix}$
Logistic (DTW)	213 (84.2%)	40 (15.8%)	0.2183	0.3315	$\begin{bmatrix} 135 & 15 \\ 25 & 78 \end{bmatrix}$
Naïve Bayes (means of attributes)	163 (81.5%)	37 (18.5%)	0.1903	0.4903	$\begin{bmatrix} 94 & 6 \\ 31 & 69 \end{bmatrix}$
Naïve Bayes (DTW)	199 (78.7%)	54 (21.3%)	0.2118	0.4408	$\begin{bmatrix} 140 & 10 \\ 44 & 59 \end{bmatrix}$
SMO (means of attributes)	166 (83%)	34 (17%)	0.17	0.4123	$\begin{bmatrix} 100 & 0 \\ 34 & 66 \end{bmatrix}$
SMO (DTW)	200 (79.1%)	53 (21%)	0.2095	0.4577	$\begin{bmatrix} 150 & 0 \\ 53 & 50 \end{bmatrix}$
Perceptron (means of attributes)	163 (81.5%)	37 (18.5%)	0.1765	0.2987	$\begin{bmatrix} 74 & 26 \\ 11 & 89 \end{bmatrix}$
Perceptron (DTW)	208 (82.2%)	45 (17.8%)	0.2086	0.3354	$\begin{bmatrix} 136 & 14 \\ 31 & 72 \end{bmatrix}$

Table 1. A summary of cross-validation results. Here, RMS error is the root-mean-squared error, and all confusion matrices in this table are of the form $A = (a_{ij})$, where a_{ij} is the number of instances classified as $y = j$ during cross-validation that were actually from group i.

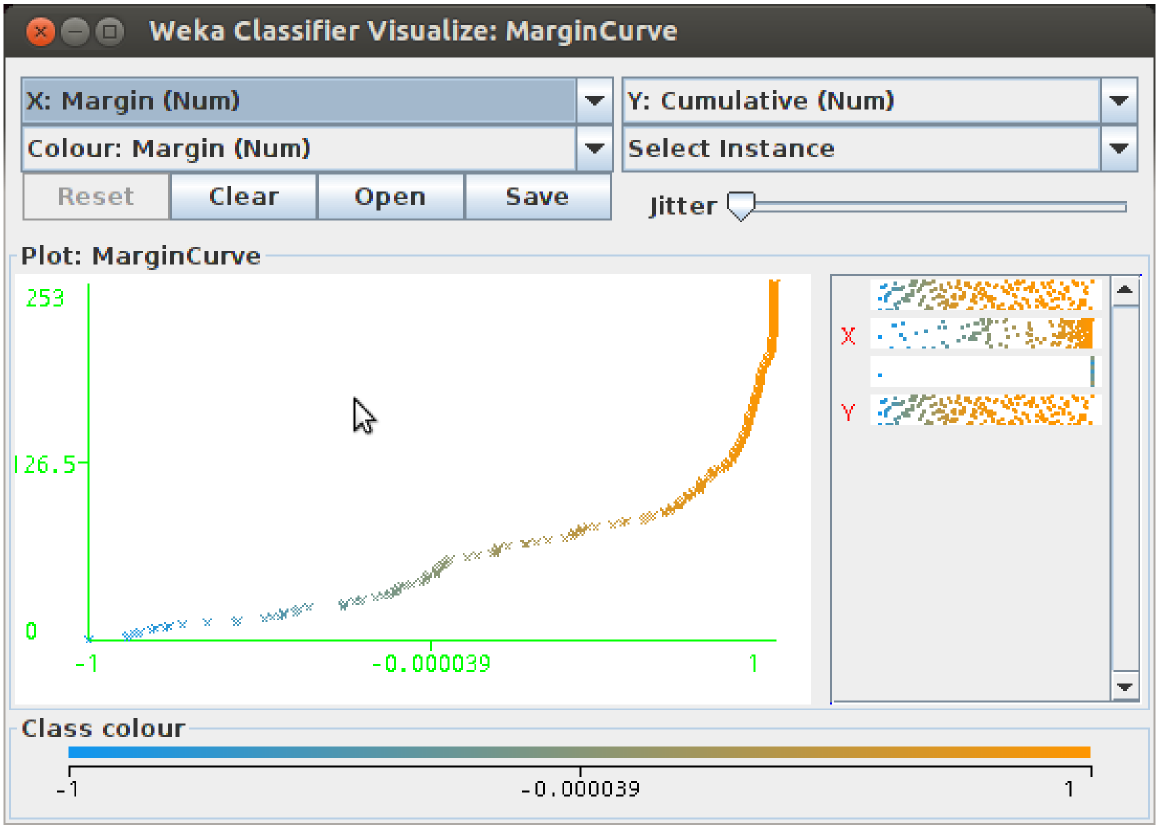


FIGURE 2. The margin curve for the logistic regression classifier. Here, the steeper the boundary line, the greater the margin between the two sets. Compare to Figure 3.

model and multilayer perceptron model compare. The multilayer perceptron model takes all the input values (the features we collected) and returns a single value between -1 and 1 that is used to distinguish light curves with exoplanets (those with value greater than 0) and light curves without exoplanets (those with value less than 0). We can look at what values our models gave for our training set. The results are plotted in Figures 2 and 3, and illustrate how the similar accuracies in these algorithms belies very different classfication boundaries.

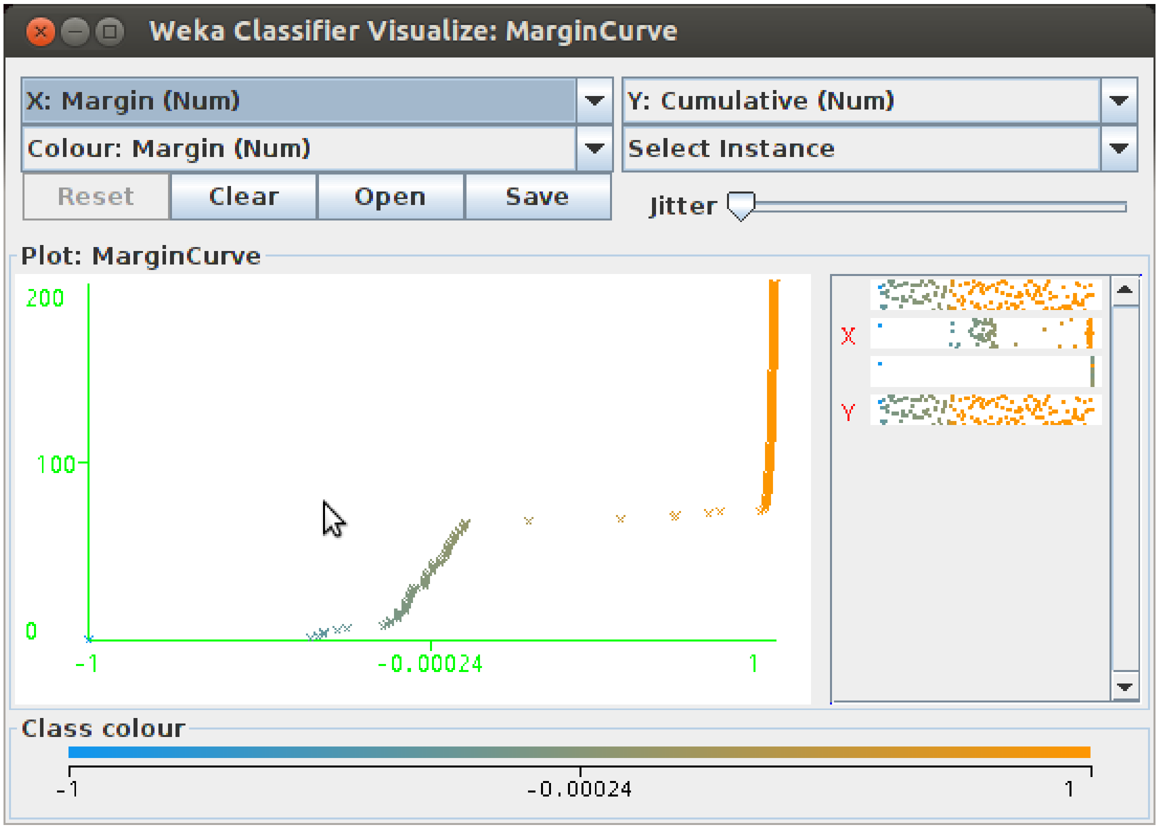
4. Analysis

We were pleasantly surprised to see such high accuracy rates from these algorithms; we went into this project assuming that issues such as noise in the data and the difficulty of programatically distinguishing transiting planets from eclipsing binaries would make this project difficult. In particular, there are good reasons to assume that the problem is not necessarily linearly separable: intuitively, there are two major astronomical constraints on a lightcurve. If there are no dips in the lightcurve, then there is probably no transiting or eclipsing body around the parent star (or it is too small to be detected with current technology); yet if the changes in the magnitude are too large, then the transiting object is unlikely to be a planet; instead, it could be a brown dwarf or even a stellar companion (in which case the system is an eclipsing binary). Thus, the exoplanets lie in between these two constraints in the "direction" of the depth of the magnitude loss (i.e. some line in the feature space that accounts for these particular differences in the lightcurves), so a linear classifier that doesn't use a higher-dimensional kernel will be inaccurate.

This could very well be why the SVMs were snarled by this classification question; their attempts to fit a clean decision boundary couldn't account for all of the edge cases. However, other classifiers are less concerned with the shape of the decision boundary, as classification is done less geometrically. Thus, these classifiers were more successful, and with feature selection optimized for time-series data, were able to perform very well.

5. Future Work

There are still several improvements that could be made. For one, we can run a clustering algorithm beforehand to group together light curves that have similar brightness and similar metafeatures. Thus, we would have clusters of stars that have similar apparent brightness. Then we can select from each cluster a light curve that contains a star that is known to not have any noticeable variability (exoplanet, binary system, variable star, etc). Then we can



Weka Classifier Visualize: MarginCurve

X: Margin (Num)

Y: Cumulative (Num)

Colour: Margin (Num)

FIGURE 3. The margin curve for the multilayer perceptron; compare with Figure 2. Even though they had very similar error rates, their margin curves look very different.

run the DTW algorithm to compare each star in that cluster to the non-variable star in order to obtain a even more "normalized" measure of similarity between light curves.

Another improvement we could make is to try to use a multinomial classifier or a clusterer to specifically classify the light curve as a binary star, a variable star, a transiting planet, or something else. Instead of just classifying to be able to tell whether there is a exoplanet or not, we could expand our attribute list and collect more features to generate a classifier that can provide more detailed classification.

We could also train and test our classifiers on larger sets of data. As the Kepler project has been discontinued, we have a very limited training set to work with. Kepler has discovered over 3000 unconfirmed planet candidates, but only a little over 100 have been confirmed, which gives us a relatively small training set. And given that greater than expected noise and failure of Kepler's instrumentation has plagued the project, the data collected in the latter half of its mission is likely dubious. In the chance of another exoplanet-discovery project in the future, we will aim to continue training and testing our classifiers. But in the meantime, there are still hundreds of thousands of Kepler target light curves to be sorted through.

REFERENCES

[1] The Extrasolar Planets Encyclopedia. `http://www.exoplanet.eu/` November 15, 2013.

[2] Malatesta, Kerri. *The Transiting Exoplanets HD 209458 and TrES-1*. American Association of Variable Star Observers. June 8, 2012.

[3] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); *The WEKA Data Mining Software: An Update*; SIGKDD Explorations, Volume 11, Issue 1.

[4] Mikulski Archive for Space Telescopes (MAST). `http://archive.stsci.edu/kepler/publiclightcurves.html` January 30, 2013.

[5] Mitchell, Scott. *The Application of Machine Learning Techniques to Time-Series Data*. University of Waikato. 1995.

[6] Robiatille, Thomas P., et. al. *Astropy: A community Python package for astronomy*. Astronomy and Astrophysics, Volume 558, October 2013.

[7] S.K. Shevade, S.S. Keerthi, C. Bhattacharyya, K.R.K. Murthy: *Improvements to the SMO Algorithm for SVM Regression*. In: IEEE Transactions on Neural Networks, 1999.

[8] *Transit Tracks*. NASA. `http://kepler.nasa.gov/education/activities/transitTracks/`. June 2013.