Search for Exoplanets

Petra Brčić
Applied Mathematics
petrabrcic94@gmail.com

Sandro Lovnički
Computer Science
lovnicki.sandro@gmail.com

June 29, 2018

Abstract: With the launch of Kepler space observatory in March 2009, the search for exoplanets has extensively began. There are currently 3,758 confirmed exoplanets and contrary to the early approaches of validating exoplanet candidates by hand, machine learning has begun taking its toll on the subject. We would like to investigate neural network model for detecting transiting exoplanets from Kepler observatory light-curve data collected over a period of 4 years, with over 200,000 documented stars in our galaxy, the Milky Way. On April 18, 2018, a new satellite (TESS) has been launched, specificaly designed to search for exoplanets using the transit method. After it's 2-year planned mission, complete data should be available to public and more than 20,000 new exoplanets are expected to be found.

Keywords: Kepler, light-curve, exoplanet, machine learning, neural network, TESS.

Contents

I. Introduction

In March 2009, NASA launched Kepler space observatory to discover Earth-size planets orbiting other stars in our galaxy. With most of the discoveries made after Kepler data was anounced, there are currently $3,758$ confirmed exoplanets in $2,808$ systems, with 627 systems having more than one planet. Contrary to the early approaches of validating exoplanet candidates by hand or with humanly constructed models, recent years have been fruitfull for usage of machine learning for classifing Kepler's data.

II. Methods

There are many methods for detecting an exoplanet existance, as it can be seen in [3].
For example, *gravitational lensing* method exploits the stars' gravitational fields' effect on magnification of distant background stars' light. If the star in question has a planet, its gravitational field makes a detectable contribution to the lensing effect. Over the past 10 years, just a 1000 such events have been observed because specific alignment of bodies is required. Nonetheless, 19 exoplanets have been discovered using this method, whose full list is given in [4].
It is notable to mention *radial velocity* method which uses variations in the speed with which specific star moves away or towards the Earth. Those variations are influenced by orbiting planet's gravitational pull and therefore we can also deduce the planet's mass. Up to 2012, this method was most effective for exoplanet detection whose full list is given in [5].

The most effective method today is the *transit photometry*. When a planet crosses (transits) in front of its star, the star's brightness diminishes depending on planet relative size to it. To detect those brightness variations, observer's relative position must be properly aligned with star-planet axis which is not something one can control. The probability of a random alignment producing a transit in a system with Sun-sized star and a planet at 1AU [1] from it is 0.47%. Ilustration of how those brightness variations are manifested in Kepler's data can be seen in Figure 1.

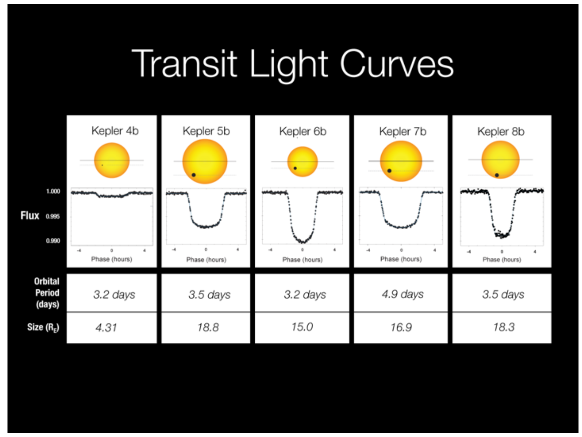

Figure 1: *Illustration from Bill Borucki's Jan 2010 AAS Presentation*

Another disadvantage of this method is a high rate of false positives due to measurement precission and external noise that can be interpreted as a transit. For this reason, transit photometry is often combined with radial velocity to produce better results and eliminate false positives. The main advantage of transit photometry is planet's size deduction from light curve drops, which combined with transit photometry's mass deduction yields planet's density - a property much appriciated in search of habitable worlds. Also, by observing spectrum of light passed through the planet's atmosphere, one can detect which chemical elements are present. By April 12, 2018, Kepler detected $2,343$ exoplanets using transit photometry as a base method.

III. Data

Data used for this project is obtained by combining NASA Exoplanet Archive data for Threshold-Crossing Events which (currently) consists of 34032 documented TCEs, each having $-CHECKNUMBER-$ column features

[1] 1 Astronomical unit $\approx$ Earth's ditstance to the Sun $\approx$ 150 million kilometres

and raw Kepler light-curves (star's flux (brightness) over time) at Mikulski Archive for Space Telescopes consisting of over $2000000 - CHECKNUMBER-$ star lightcurves monitored by Kepler.

Most important TCE features for our projects are `kepid`, `tce_period`, `tce_time0bk`, `tce_duration` and `av_training_set` which (respectively) correspond to the Kepler identifier of host star, period of the event, time of the first appearance (more precisely; time of the begining of drop + `tce_duration`/2), duration of the event (drop in star's brightness) and label from set $\{PC, NTP, AFP, UNK\}$ meaning planet candidate, non-transit phenomena, astrophisical false positive and unknown. Those features will enable as to download just the Kepler light-curves of interest and later on to process each light-curve to extract and amplify features of TCE.

i. Creating Dataset

At the time of collecting our data, an older version of a TCE table was available with around $20000 - -GETCORRECT-$ labeled TCEs. Distribution of `av_training_set` labels is as follows:

- -NUMBER- TCEs with label PC (planet candidate)
- -NUMBER- TCEs with label NTP (non-transit phenomena)
- -NUMBER- TCEs with label AFP (astrophisical false positive)
- -NUMBER- TCEs with label UNK (unknown)

We used that table, specifically `kepid` colum to download just the light-curves for those stars from Kepler data. These light-curves are taking about 90GB of space and are just a fraction of all the Kepler data.

Still, our machines are not fully prepared to tackle the preprocessing and learning from such a vast amount of not at all simple data, so we decided to choose randomly just some of the TCEs from table. We ended up selecting 1058 TCEs labeled as planet candidate, 471 TCEs labeled as astrophisical false positive and 734 TCEs labeled as non transit phenomena.

As an example, here is a small table of selected features of one TCE documented for a star with `kepid` 9517393:

kepid	9517393
tce_period	219.322
tce_time0bk	320.728
tce_duration	0.5125000000000001
av_training_set	PC
...	...
tce_plnt_num	1
tce_eqt	313.0
...	...

ii. Preprocessing Lightcurves

In order to feed our model, we first need to process the raw light-curve into more useful and compact form. Raw light-curve is loaded as an array of smaller arrays called quarters which represent distinct periods in Kelper telescope's configuration, position, etc. We thus might expect quarters to be on a slightly different scale. In figure 2 we show lightcurve for `kepid` 9517393 right after loading `time` and `flux` arrays.

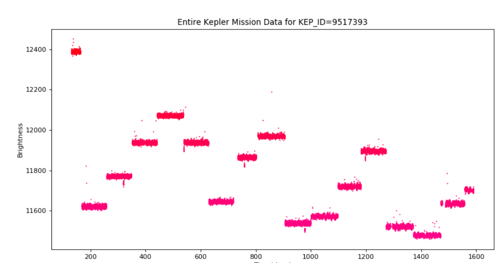

Figure 2: *Raw light-curve data*

It is easy to see that almost every star has an unpredictable variation in its flux, so the data is pretty distorted and irregular which is something that is to be avoided because it would interfere with our model's conclusions and feature extraction. Thus, we calculate a B-spline over the entire light-curve data, ignoring points that cannot be fitted or are of type None. This spline can be seen in figure!3 colored black.

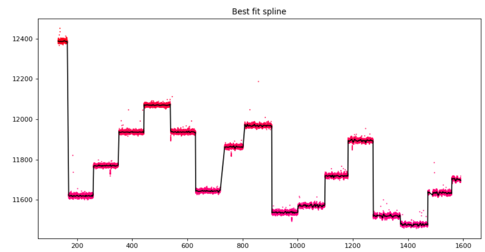

Figure 3: *Raw light-curve data with fitted B-spline*

Next, we divide light-curve by the spline to obtain 4. Notice how specific TCEs can now be seen without even enlarging.

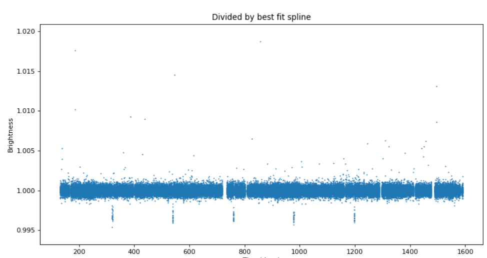

Figure 4: *Raw light-curve data divided by spline*

Now we need to exactly find the position of desired TCE which we know from TCE table, column `tce_time0bk`. Focusing just on that segment of this plot, we get figure 5

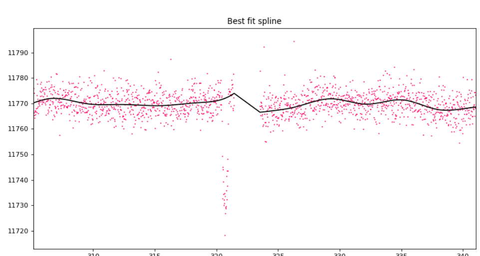

Figure 5: *Enlarged area of the TCE*

We see that the amount of useful data of this "drop" is very poor, but we must remember that those events are periodic by definition so we hope that there is more instances of that same "planet" crossing during the lifetime of Kepler. We find the `tce_period` from TCE table and fold every drop separated by `tce_period` to `time0bk`. Now we divide TCE's duration into 201 discrete points and calculate median value for flux in each of 200 intervals in between. This results in a much richer, and uniform for all events, representation of a TCE which can be seen in!6. We call that view a "local view" and it is going to be the input of our model.

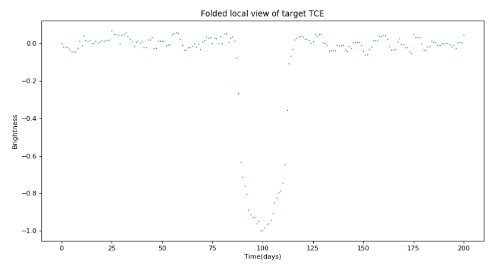

Figure 6: *Folded local view of the TCE*

IV. Convolutional Neural Network

i. 1D Convolution

ii. ReLu

iii. Pooling

V. Training

i. Parameters

ii. Progress

VI. Results

References

[1] `https://en.wikipedia.org/wiki/Kepler_(spacecraft)`

[2] `https://en.wikipedia.org/wiki/Transiting_Exoplanet_Survey_Satellite`

[TESS] https://tess.mit.edu/

[3] https://en.wikipedia.org/wiki/Methods_of_detecting_exoplanets

[4] https://en.wikipedia.org/wiki/List_of_exoplanets_detected_by_microlensing

[5] https://en.wikipedia.org/wiki/List_of_exoplanets_detected_by_radial_velocity

[6] https://en.wikipedia.org/wiki/List_of_exoplanets_discovered_using_the_Kepler_spacecraft

[Shallue and Vanderburg, 2017] Christopher J. Shallue, Andrew Vanderburg (2017). Identifying Exoplanets with Deep Learning: A Five Planet Resonant Chain around Kepler-80 and an Eighth Planet around Kepler-90 arXiv:1712.05044 [astro-ph.EP]

[Mikulski Archive for Space Telescopes] https://archive.stsci.edu/

[NASA Exoplanet Archive] https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=q1_q17_dr24_tce

[TCEs table] https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=q1_q17_dr24_tce

2