Refining Exoplanet Detection Using Supervised Learning and Feature Engineering

SLIOIA

M. Bugueño, F. Mena and M. Araya

Departamento de Informática
Universidad Técnica Federico Santa María

October 2018

CLEI LACLO 2018

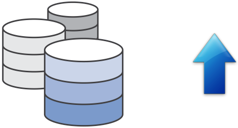
EX UMBRA IN SOLEM

Summary

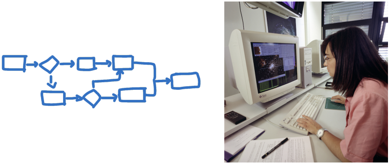
1. Introduction
2. Background
   - State of the art
3. Data
   - What and Where
   - Inside the data
4. Models and Methods
   - Feature extraction
   - Learning models
   - Evaluation
5. Experiments
   - Results
   - Detail of results
   - Tagging
6. Conclusions

Motivation

Growing data



Automate processes



A lot of manual process and analysis

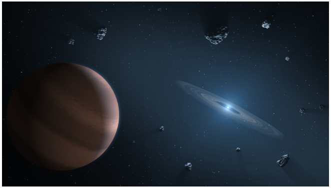
Differents fields of application

What is an exoplanet ?

Planets orbiting stars outside our solar systems are called **extra-solar planets** or **exoplanets**.

Detecting these planets is a challenging problem !

- They emit or reflect very dim magnitudes compared to their host stars
- They are very near to their host stars compared to the observation distance
- Fine-grained analysis is needed

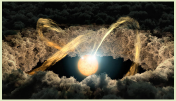


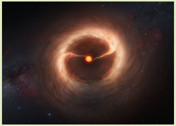
Why is difficult to detect exoplanets ?

The theory of extrasolar planets is under development until these days (since the first confirmed detection of the giant planet **51 Pegasi b**)

Birth of a planet

- Current theories suggest that the dust particles of the protoplanetary disk begin to collapse by gravity forming larger grains.
- If these discs survive to stellar radiation and comets or meteorites, the matter continues compacting giving way to a planetoid.
- Unfortunately, due to limitations in detection methods, most of the discovered planets are the big ones.





Images by [1] and [2]

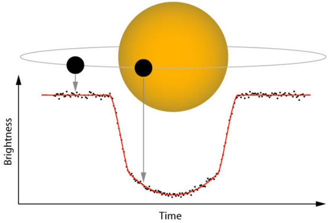
[1] http ://www.spitzer.caltech.edu/news/1876-feature16-07-Light-Echoes-Give-Clues-to-Protoplanetary-Disk

[2] https ://www.space.com/19100-alien-planet-birth-alma-telescope.html

Detection methods

The most successful detection mechanisms are the **indirect**, some of them :

- ♂ **Radial velocity**, which studies the speed variations of a star product of its orbiting planets, analyzing the spectral lines of this one through the Doppler effect. *Successful, but only effective on giant planets near its star*
- ♈ **Transit photometry**, photometric observation of the star and detection of variations in the light intensity when an orbiting planet passes in front of it. *Efficient, detect high-volume planets independently of the proximity to its star*

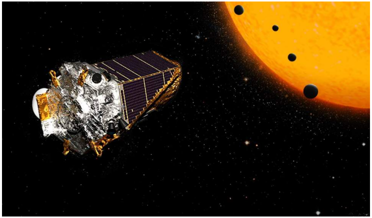


What is needed ?

Problem : A large volume of data is being generated today !

- The use of automated methods could reproduce the astronomer analysis to decide if the data supports the existence of an exoplanet or not.

Fortunately, technological advances in photometry have allowed experiments like the space observatory Kepler to have sufficient sensitivity for detecting a greater range of exoplanets.



Previous work

- Richards et al. (2011) presents a catalog of variable stars and manually extracted light curve specialized features from simple statistics and other features based on the period and frequency analysis of a LombScargle fitted model
- Donalek et al. (2013) classify variable stars from the Catalina Real-Time Transient Survey (CRTS) and the Kepler mission extracting similar features from the light curve to Richards et al.
- Mahabal et al. (2017) also with the propose of classify variable stars, transformed light curves into an image (grid) that represent the variations of magnitude through the variations of time intervals.
- Hinners et al. (2017) presents different machine learning techniques and models with the objective of classify and predict features over the same data as we use. They extract some statistical features from the light curve but they were not interested on exoplanet detection.

Our Objective : Exoplanet detection with ad-hoc feature extraction and machine learning.

Where the data come from ?

Kepler space observatory

- Data collected by Kepler Mission (launched in 2009 by NASA) with the goal of searching similar planet to Earth.
  - Around 65% of exoplanet discoveries have been detected thanks to Kepler Mission [3]
  - As most of the discovered exoplanets have been detected trough the transit method, and taking advance of photometric improvements of Kepler.
- The Kepler Objects of Interest (**KOI** [4]) dataset is provided by MAST (Mikulski Archive for Space Telescopes).
  - It containts 8054 KOI's.

[3] https ://exoplanetarchive.ipac.caltech.edu/docs/counts_details.html

[4] http ://archive.stsci.edu/search_fields.php ?mission=kepler koi

What are the data ?

The Labels

- Every record is associated to a Kepler Object of Interest labeled as :
  - *Confirmed* (2281) : those that have been confirmed as exoplanet, through extensive analysis.
  - *False Positive* (3976) : those that were initially selected as candidate exoplanets but there is additional evidence that shows they are not.
  - *Candidate* (1797) : those that are still under study.

  according to Nasa Exoplanet Science Institute[5]
- The reasons to catalog as a False Positive are observation that did not match with the star position on study. Also can be that the deep of the even transit was statistically different to the deep of the odd transits, showing a binary system.

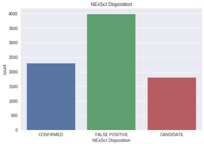


FIGURE – categories distributions on collected data

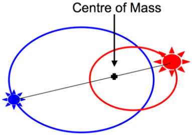


FIGURE – binary star system

[5] https ://exoplanets.nasa.gov/resources/269/keplers-greatest-hits

What contain each KOI

Error

- The error associated to each measure

Time

- The time (in Julian Date - January 1, 4713 BC) when the measure was made

Raw light curve

- The raw measurements (intensity) of star light, with about 70000 measurements
- The series is governed by a trend and could have cycles.
  - This fact is important because the Kepler measurements are not recorded uniformly.
- On average, the missing data is about 23% , this mean approximately about 55000 *effective* measurements. We used two simple techniques to tackle this :
  1. fill missing with zeros
  2. fill the gaps with linear interpolation

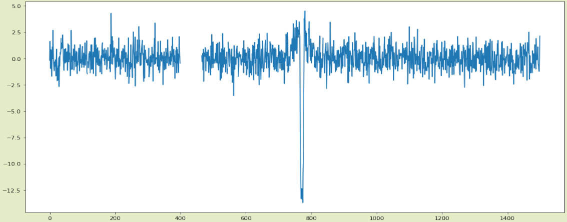
What contain each KOI

Filtered light curve

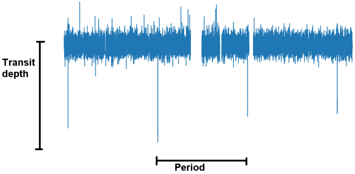
- Light curve with a **whitened filter**. This transform led a constant white noise on the light curve, i.e., the higher signal get amplified and give a more uniform signal.
- The white noise is a random signal that have the same intensity on different frequencies, which gives a spectral density of constant power. The operation is as follow :

$$lc_{whitened} = \frac{lc_{raw}}{S_{lc_{raw}}(f)} \tag{1}$$

S is the spectral power density and f is the frequency.



What contain each KOI - Metadata



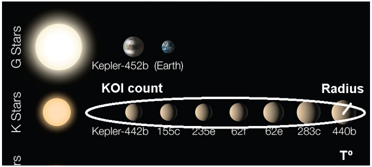


FIGURE – planet metadata, right image by [6]

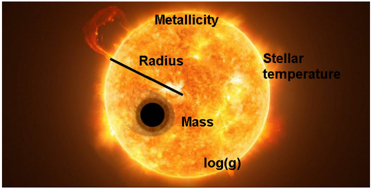


FIGURE – stellar metadata, image by [7]

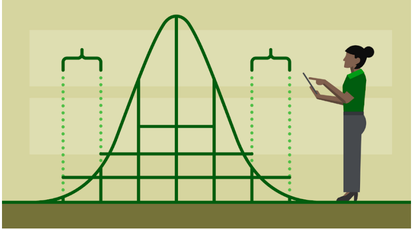
[6] https ://exoplanets.nasa.gov/resources/269/keplers-greatest-hits

[7] https ://www.tititudorancea.com/z/metals_signs_water_found_exoplanet.htm

Manual

We used extraction techniques specialized on time series, which in this case corresponds to measurements of intensity of the light along time, inspired on the library Feature Analysis for Time Series (**FATS**[8]) for Python.

- Amplitude
- Slope
- Max
- Mean
- Median
- Median absolute deviation
- Min
- Q1
- Q2
- Q31
- Residual bright faint ratio
- Skew
- Kurtosis
- Std



To this we decide to add the metadata

[8] https ://github.com/isadoranun/FATS

Automatic

- We use : **Unsupervised Learning Methods**
  the objective is to find intrinsic patterns among all the data independently from task.

Pre-process

- Firstly we applied a Discrete Fourier Transform[1] to the light curve.
  - It transforms the data from the time domain in which the measurements where obtained, to the frequency domain where the signal was generated.

[1] *this method is designed to analyze periodic signals, which is exactly the case of transit light curves.*

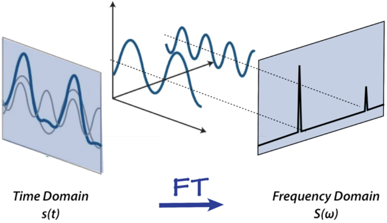


FIGURE – image by http ://mriquestions.com/fourier-transform-ft.html

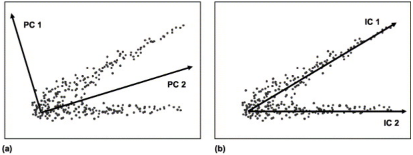
Automatic

PCA

- The **Principal Component Analysis**, as a linear method that projects data into a lower dimensional space (the higher variance vector).
- Why PCA ?
  - Has been applied to several applications, obtaining particularly good results on time series
  - Its great efficiency from specific optimizations over linear algebra methods with high dimensional data.

ICA

- The Fast iterative algorithm for **Independent Component Analysis** that finds statistically independent components of the data.
- Why ICA ?
  - Is focused on the signal abstraction, since it tries to detect the independently sources that, mixed, produce the observed data.
  - Differs to the uncorrelated componentes that finds PCA.



Machine Learning models

k-NN, regularization parameter : *k*

- Based on memory (i.e., non-parametric). Simple, but with good performance

Logistic Regression, regularization parameter : *C*

- Classify based on a probabilistic (logistic - sigmoid function) linear model.

SVM, regularization parameter : *C*

- A linear margin-based model... but with Kernel : RBF (Radial Basis Function)

Random Forest, regularization parameters : *depth* and *T*

- A ensemble of decision trees.. not linear !

Neural Network, recurrent with gates : LSTM and GRU

- A non-linear composition model over statistics by windows representation.

Metrics

Classification

The exoplanet problem is an instance of **unbalanced binary** classification problem.

- Precision : ability to label one class when the object effectively was from that class. (inverse to contamination)

$$P = \frac{T_p}{T_p + F_p}$$

- Recall : ability to include all object that effectiveley are from one class. (similar to completeness)

$$R = \frac{T_p}{T_p + F_n}$$

- F_1-score : as harmonic mean between P and R

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R}$$

All these metrics reach their best values at 1 and worst at 0

Experiments

- Large amount of data to be processed ! So it was necessary to use a cluster provided by **ChiVO** [9] (Chilean Virtual Observatory)
- We have 6257 labeled data, corresponding to 121*GB*, and 1797 not labeled data (candidates), corresponding to another 33*GB*.
  - We used the (64/18/18)% of the labeled data for set-split. Approximately 4000, 1000 and 1000 registers were grouped as training, validation and testing sets. This last one represents the actual *target* unlabeled data that is unknown (Candidates).

Note : The selection of the hyper-parameters of the classifiers was not expensive in computational terms (dimensionality *d* much smaller than the original), unlike the feature extraction process.

[9] http ://www.chivo.cl

Experiments

- For automatic feature extraction techniques, fixed dimensions were experimented

	5	**10**	**15**	**20**	**25**	**50**
ICA	**0.711**	0.709	0.709	0.686	0.679	0.675
	5	**10**	**25**	**55**	**100**	**255**
PCA	**0.713**	0.701	0.701	0.699	0.702	0.689

TABLE – F1 score of the best classifier, Random Forest, in function of dimensionality.

- Surprisingly enough, completing missing data with zeros produces a consistent improvement of ~ 0.1 in the F_1-score, while linear interpolation produced worse.
- Another technique consists in performing a sampling of the sequence by taking the maximum value each 3 points (considering the missing data as zeros) completing the data following the trend line. Unfortunately, this resulted in a greater error.

What about unbalanced data ?

- **Undersampling technique** ; the majority class was subsampled
  - ..but **Unbalanced data** over Logistic Regression, SVM and Random Forest models
- Some models admit **weighting** different classes
- Over this last experiment we improved the results by ~ 0.1 on F_1 score metric.

Performance Results

Learners performance

	k-NN	*Logistic Regression*	*SVM RBF*	*Random Forest*
Fourier + PCA	0.679	0.493	0.486	**0.713**
Fourier + ICA	0.679	0.493	0.486	**0.711**
OwnFATS	**0.666**	0.583	0.575	0.658
Planet metadata	0.825	0.848	0.848	**0.870**
Stellar metadata	0.766	0.718	0.751	**0.766**
OwnFATS + stellar & planet metadata	0.844	0.864	0.876	**0.883**

Table – F1 score on the classification of different models (learners) over the test set on the different representations generated.

Results analysis

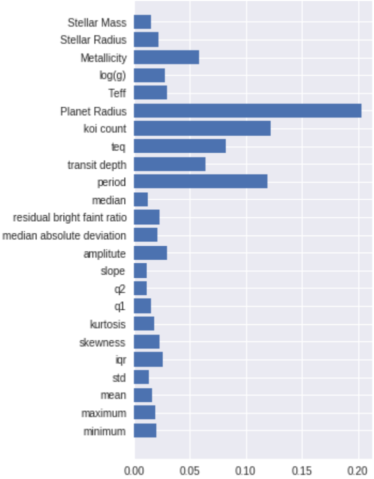


FIGURE – Random Forest feature importance

Performance analysis

- The best trained model, was Random Forest
- Performance of 88.3% for future classification by F_1 score metric
- In detail :
  - Radius, period and number of objects in the system as the most relevant features
  - The less important features are the extracted from the light curve (slope and second quartile)

EX UMBRA IN SOLEM

Precision-Recall detail

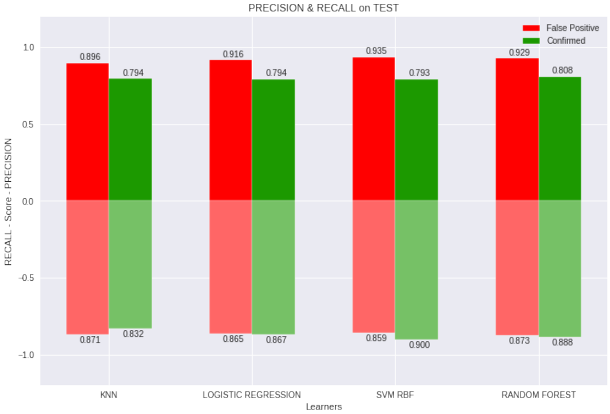


FIGURE – Manual extraction features plus metadata

- The best model that can identify correctly the False Positive Class (FPC) based on Precision and Recall is the SVM RBF.
- Classification on the Confirmed class shows lower scores than FPC suggesting the difficulty on the exoplanet prediction.
  - Why ? Could be because all of them do not have very similar features on the light curve
- The best model in the task of confirmed exoplanets on the different representation of the data was Random Forest.

Final Results

Selected representation :

- OwnFATS + stellar & planet metadata

Selected model for *confirmed* class :

- Random Forest ($depth = 15$, $T = 15$)

Selected model for *false positive* class :

- SVM RBF ($C = 100$)

We show the classification over the Kepler Object of Interest that are still being studied by the staff of NexScl on September 2017 :

Total CANDIDATE	1791
Subtotal CONFIRMED	975
Subtotal FALSE POSITIVE	434
Unclassified	382

KOI name	**Disposition**	**Confirmed on system**	**Star**
K00601.02	*FALSE POSITIVE*	2/3	Kepler 619
K00750.02	*UNCLASSIFIED*	1/3	Kepler 662
K01082.01	*CONFIRMED*	1/4	Kepler 763
K01082.02	*FALSE POSITIVE*		
K01082.04	*CONFIRMED*		
K01236.04	*CONFIRMED*	2/3	Kepler 279
K01358.01	*CONFIRMED*	0/4	-
K01358.02	*CONFIRMED*		
K01358.03	*CONFIRMED*		
K01358.04	*CONFIRMED*		
K01750.02	*CONFIRMED*	1/2	Kepler 948
K02064.01	*UNCLASSIFIED*	0/1	-
K02420.02	*CONFIRMED*	1/2	Kepler 1231
K02578.01	*FALSE POSITIVE*	0/1	-
K02828.02	*FALSE POSITIVE*	1/2	Kepler 1259
K03444.03	*UNCLASSIFIED*	0/4	-
K03451.01	*UNCLASSIFIED*	0/1	-
K04591.01	*FALSE POSITIVE*	0/1	-
K05353.01	*FALSE POSITIVE*	0/1	-
K06267.01	*CONFIRMED*	0/1	-

github.com/FMena14/ExoplanetDetection

Conclusions

- We introduced a new refining method to decide if an object on study (KOI) is really an exoplanet using automatic learning and handling raw data
  - We reproduce the arduous and extensive work that experts perform on detection
- The results show that the automatic techniques used to extract information from the light curve was not good enough compared to the metadata
  - Maybe the not suitable methods for the feature extraction, or well, too simple for the complex problem that we faced
- Also the problem was complex regarding the execution time **on feature extraction**
- The decision about which metadata were used could have great impact on the results informed in this work

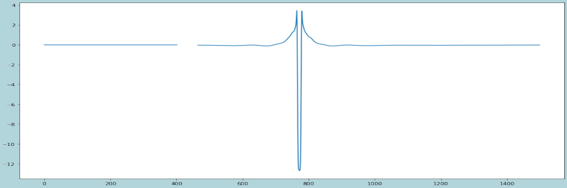
Acknowledgments : Thanks to Chilean Virtual Observatory, ChiVO. Also we thanks to the academic Ricardo Ñanculef.

Working on...

Future Work

Mandel-Agol fit

- The Mandel-Agol model the transit of a stratospheric planet around a stratoshperic star, like an eclipse, assuming a uniform source.
  - It requires the distance from the planet to the parent star as well as the radius of each one of the bodies.
- The closeness of the planet to star (eclipse) is modeled as a quadratic polynomial.



- Also is the modification of the fill in techniques on the missing values or some techniques that can handle this properly.

Refining Exoplanet Detection Using Supervised Learning and Feature Engineering

SLIOIA

M. Bugueño, F. Mena and M. Araya

Departamento de Informática
Universidad Técnica Federico Santa María

October 2018

EX UMBRA IN SOLEM

What about recurrent neural network ?

- We tried with different representations on the input data :
    - Size of the window (300-500)
    - Size of the stride (100-150)
- No good results were achieved for any the networks, even varied hardly the architecture of the network
- A little network with GRU's gate was the one with the best performance, 0.567 according to F_1 score.

We suspect that the sequence was too long so the statistic by window was not the best technique to summarize information for the proper learning of the network