Can we interpret machine learning? An analysis of exoplanet detection problem

Gabriel Molina, Francisco Mena, and Margarita Bugueño

Universidad Técnica Federico Santa María, Santiago, San Joaquín, Chile;
`gabriel.molina.12@sansano.usm.cl`

Abstract. The exoplanet detection problem - planets that orbit a star outside our Solar System - has focused on the use of time-consuming manual process. Now, the promising techniques are machine learning methods. However, the lack of interpretability in order to understand what the models does, has avoided the improvement and development of the models. In this work, we study the use of classical machine learning methods for detecting confirmed objects on the Kepler mission. Using metadata from the objects and hand-crafted features from the *light curves*, our study shows that approximately 93% of the data is correctly detected. The extreme behavior of non-exoplanet objects facilitate the recovery of mostly all these objects (high *recall*), however our work presents difficulties with confirmed objects overlapping with the non-exoplanet objects (low *precision*). Because of this, we provide some insights about where the error could be in order to interpret the learning process of our proposal.

1. Introduction

The detection of exoplanets in the large amount of data generated by the observatories is an urgent issue. In order to study the possibilities of life outside the Earth even discover another systems. The exoplanets emit or reflect very dim magnitudes of light compared to their host star, and they are very near to them compared to the observation distance, hindering the discovery. Therefore the exoplanet detection problem is a challenging task since it could be faced using different methods and approaches. Currently the most successful techniques carried fine-grained and time-consuming analysis of some indirect methods, where the transit method - analysis of periodicities in photometric observation of a start - is a pioneer. Today, some of these methods have being extended to automatic techniques (McCauliff et al. 2015), as machine learning, that could reproduce in less time the astronomer analysis. Despite of the good performance of the models they have some limitations, keeping the progress stagnated in different aspects, i.e. they lack of interpretability on what the model learns and does.

This paper presents a study on exoplanet detection problem using the technological advances on the Kepler mission and ML methods on *light curves* measurements. This work also presents an analysis of the interpretability of the model predictions understanding why it fails in some data. We compare our known solar system planets to the interpretability to add references points in order to clarify the analysis.

This paper is organized as follow: 2) present the data that we used, 3) comments the models and representation explored in the classification, 4) shows the results with a brief analysis of these. Finally, in 5) the conclusions are commented.

2. Data

The photometric observation of some star generates a time series that has some variations depending on the intensity of the light, called *light curve*. When a orbiting planet passes in front of this star, blocking a fraction of the light is called a *transit*. In the present work we focus on the transit method for exoplanet detection.
Given the documentation and effectiveness of the mission, our work use the Kepler[1] dataset. This is the largest labeled dataset on exoplanet detection which is composed of 9564 Kepler Object of Interest (KOI). A KOI - candidate object - can be confirmed as exoplanet, rejected based on additional evidence or still be under study (unlabeled). We use the objects that have their light curves available i.e. 2281 *Confirmed*, 3976 *False Positive* and 1797 *Candidate*. In addition, each KOI has features that we select related to the study itself, called metadata. Some of them were the period of the transit, orbit radius, metallicity and temperature of the host star, the object and stellar radius, among others reporting a total of 58 features.

3. Models and Methods

Based on the type of the data, we focus on generate different representations as input for a Random Forest model (RFM) to learn the class mapping, as previous work used (McCauliff et al. 2015; Hinners et al. 2018; Bugueno et al. 2018).
Knowing some metadata of the transit object, a light curve fit can be done to get a smooth version of it, like the Mandel-Agol model (Mandel & Agol 2002). We applied a *Discrete Fourier Transform* (DFT) to this light curve and extract features with two techniques: *PCA* and *ICA*, obtaining two automatic reduced representation of the long *light curves*. Also we experiment with hand-crafted techniques, i.e. extracting manual features of the *light curves* based on summary statistics, as previous works have applied on variable star detection (Richards et al. 2011; Donalek et al. 2013). This features were concatenate to the metadata, obtaining 72 manual features to train the RFM.
In order to reduce and simplify the manual features, we generate a new representation based on a feature selection technique called FSS (*Forward Stepwise Selection*) (Caruana & Freitag 1994). This method incrementally select and add one feature at a time, starting empty, based on the improvement with respect to some performance metric.

4. Results & Discussion

As (Bugueno et al. 2018) shows, the exoplanet detection is a unbalanced binary classification that needs correctly selected metrics to measure the performance. We use the *precision* (P), *recall* (R) and *f1 score* (F1) for each class. This last metric is also averaged at macro, micro and weighted.
We create a test set for evaluation and use cross-validation for model selection. The final results obtained in the test set by the different representation that we generated is shown in Table 1. This shows that selecting 20 features over the manual representation

[1] Kepler measured the light variation of thousand of distant stars for a period of four years, with a sampling rate of half an hour, in search of periodic planetary transit. The dataset is provided by MAST: `http://archive.stsci.edu/search_fields.php?mission=kepler_koi`

Table 1. Different evaluation metrics on the classification of test set by the Random Forest model (RFM). The results on all the generated representations are shown, where *d* represent the number of features. Also in bold is the best on each metric.

		Confirmed			*False Positive*			*Global* F1		
Representation	*d*	P	R	F1	P	R	F1	Macro	Micro	Weight
Fourier-PCA on Mandel-Agol	25	63.04	80.40	70.57	89.25	77.55	82.99	76.83	78.47	79.01
Fourier-ICA on Mandel-Agol	25	64.57	77.28	70.33	88.03	79.81	83.72	77.02	78.98	79.40
Metadata + Manual features	72	86.46	92.28	89.27	96.20	93.11	94.63	91.96	92.84	92.90
(FSS) Metadata + Manual features	20	**87.78**	**92.48**	**90.07**	**96.32**	**93.87**	**95.08**	**92.57**	**93.42**	**93.46**

"(FSS) Metada + Manual features" is quite meaningful in order to recognize patterns effectively by the RFM. This representation out-performances all the other with metrics above 87% in the *Confirmed* class, 93% in the *False Positive* class and all the globals F1 with values close to ~ 93%. The detailed metrics by class shows that the *False Positive* is easier to detect than *Confirmed* even across all the representations, having always high P and F1. Also it shows the difficult on having a good P value on *Confirmed* class, indicating that the predictions of this class are usually contaminated.

One interpretation technique that we used, linked to Random Forest, is to get a feature importance score based on the decision tree ensemble. On Figure 1 is presented the 20 most importance features, here we can see that the errors became very important for the model decision, being the Metallicity error along with Planet Radius and KOI count the ones with more impact which were also selected by the FSS method. Related to the Planet Radius feature, as the box-plot on the same Figure shows, the *False Positive* class has quite extreme values and high variability respect to *Confirmed* class, that clarify the well detection of this class by the model and the impact if the feature by itself.

We project the data using Kernel PCA and visualize the error on both classes in Figure 2. As we expected, our model fail in the region close to the decision boundaries showing that there is still some overlap among classes that could be improved with features modification. In the right side of the figure we add some well known planets of our solar system and Proxima Centauri b projecting just the features that we known of those planets, like planet radius and temperature. Our visualization analysis indicated that similar planets to Mercury has been discovery, thanks to the closeness to its star, but with quite close error, while there is a gap of discovered object on Earth-like and Jupyter-like planets - relatively far of its host star. Based of it, more confirmed objects of these kind would improve the results of the methods.

5. Conclusions

In this work we study the use of classic machine learning methods for detect confirmed objects on the Kepler mission using four different representation of the Kepler data. With the metadata in addition to hand-crafted features, we reached values close to 93% on the F1 scores. These results demonstrate the utility of hand-crafted features to improve the classification showing that we can still extract information from the raw *light curves*. In the visualization, our model fail in the region close to the decision boundaries showing that there is still some overlap among classes that could be improved.

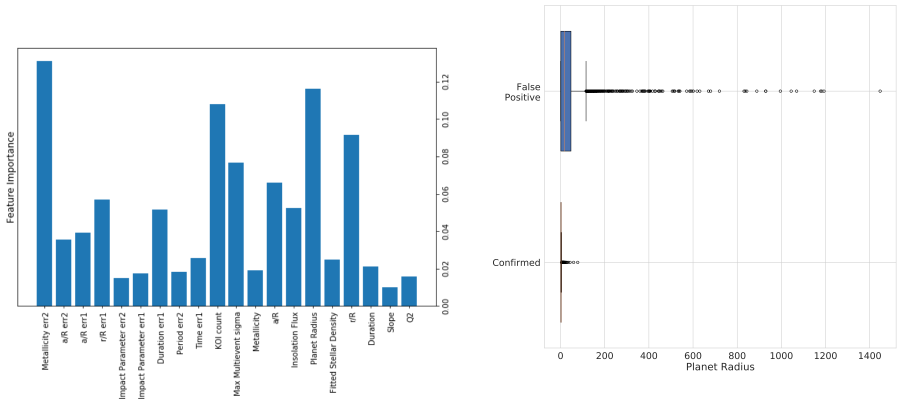

Figure 1. *Left*: Random Forest feature importance over FSS representation. *Right*: *Planet Radius* feature box-plot distribution.

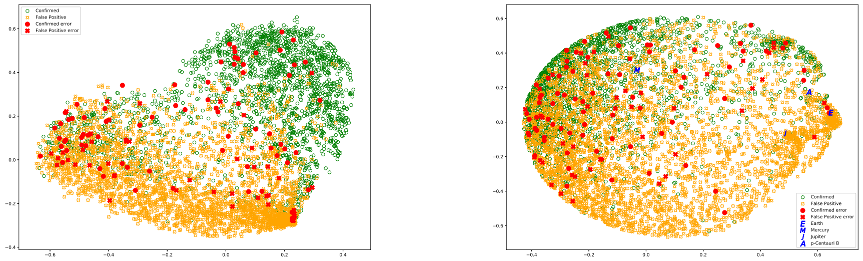

Figure 2. Projection of the data with Kernel PCA over some comparative features, colored by class, filled red symbols represent missclassified data. *Left*: Over all features. *Right*: Over some features with some extra planets.

Acknowledgments. The authors would like to thanks the *Programa de Iniciación Científica* PIIC-DGIP of the Federico Santa María University for funding this work.

References

Bugueno, M., Mena, F., & Araya, M. 2018, in 2018 XLIV Latin American Computer Conference (CLEI) (IEEE), 278

Caruana, R., & Freitag, D. 1994, in Machine Learning Proceedings 1994 (Elsevier), 28–36

Donalek, C., Djorgovski, S. G., Mahabal, A. A., Graham, M. J., Drake, A. J., Fuchs, T. J., T., M. J., et al. 2013, in Big Data, 2013 IEEE International Conference on (IEEE), 35

Hinners, T. A., Tat, K., & Thorp, R. 2018, The Astronomical Journal, 156, 7

Mandel, K., & Agol, E. 2002, The Astrophysical Journal Letters, 580, L171

McCauliff, S. D., Jenkins, J. M., Catanzarite, J., Burke, C. J., Coughlin, J. L., Twicken, J. D., Tenenbaum, P., Seader, S., Li, J., & Cote, M. 2015, The Astrophysical Journal, 806, 6

Richards, J. W., Starr, D. L., Butler, N. R., Bloom, J. S., Brewer, J. M., Crellin-Quick, A., Higgins, J., Kennedy, R., & Rischard, M. 2011, The Astrophysical Journal, 733, 10