Refinamiento en la detección de Exoplanetas mediante *Machine Learning* y *Feature Engineering*

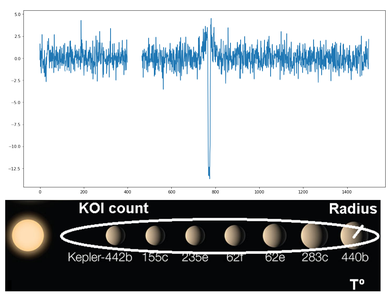
Margarita Bugueño y Francisco Mena, Mauricio Araya
Departamento de Informática, Universidad Técnica Federico Santa María

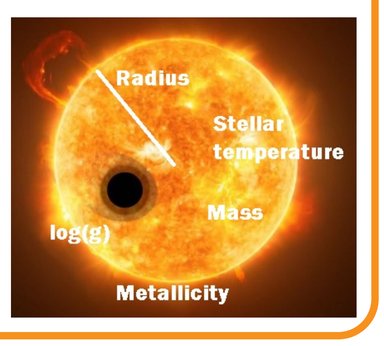
1. Resumen

El análisis de datos astronómicos ha experimentado un importante cambio de paradigma en los últimos años. La automatización de ciertos procedimientos ya no es una característica deseable sino que es necesaria para hacer frente a los muchos *datasets* extremadamente grandes que se producen hoy en día; En particular, para la detección de exoplanetas. Saber si la variación dentro de una curva de luz es evidencia de un planeta, requiere aplicar métodos avanzados de reconocimiento de patrones a un gran número de estrellas candidatas. Por esto, presentamos un enfoque de aprendizaje supervisado para refinar los resultados producidos por el análisis caso-caso de curvas de luz, aprovechando el poder de generalización de las técnicas de aprendizaje automático para predecir las curvas de luz no clasificadas a la fecha. Así, mostramos que esta técnica puede acelerar el costoso proceso manual que actualmente realizan los científicos expertos.

2. Datos

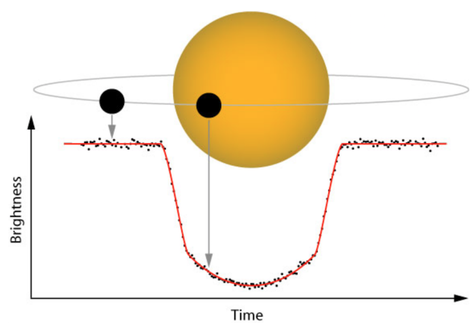
- Kepler Objects of Interest (***KOI***) [1], Misión Kepler de la NASA.
- Tres tipos de objetos: Confirmados (2281), Falsos Positivos (3976) y Candidatos (1797)
- Se trabaja con curvas de luz y metadata (planeta y estrella)





3. Métodos y Modelos

Se trabajó con el método de tránsito fotométrico



Generación de características:

- Extracción Manual de atributos basados en FATS [3], como trabajos previos [2].
- Extracción automática post Fourier (PCA vs ICA)

Se utilizaron modelos de aprendizaje clásicos con buen desempeño, como es k-NN, SVM con kernel RBF, Regresión Logística y Random Forest. Debido a la data faltante de las observaciones Kepler, se utilizaron dos técnicas simples de rellenado: autocompletar con ceros e interpolación lineal.

8. Referencias

[1] Kepler KOI. `archive.stsci.edu/kepler/koi/search.php`, Enero 2013.

[2] T. Hinners, K. Tat, and R. Thorp. Machine learning techniques for stellar light curve classification. *arXiv preprint arXiv:1710.06804*, 2017.

[3] I. Nun, P. Protopapas, B. Sim, M. Zhu, R. Dave, N. Castro, and K. Pichara. Fats: Feature analysis for time series. *arXiv preprint arXiv:1506.00010*, 2015.

[4] M. Solar, M. Araya, L. Arévalo, V. Parada, R. Contreras, and D. Mardones. Chilean virtual observatory. In *Computing Conference (CLEI), 2015 Latin American*, pages 1–7. IEEE, 2015.

4. Experimentos y Resultados

Los datos trabajados (150 GB) fueron procesados en el cluster de ChiVO (*Chilean Virtual Observatory*) [4], contando con 6257 datos etiquetados divididos como entrenamiento, validación y pruebas (64/18/18)%, además de otros 1797 no etiquetados (candidatos).

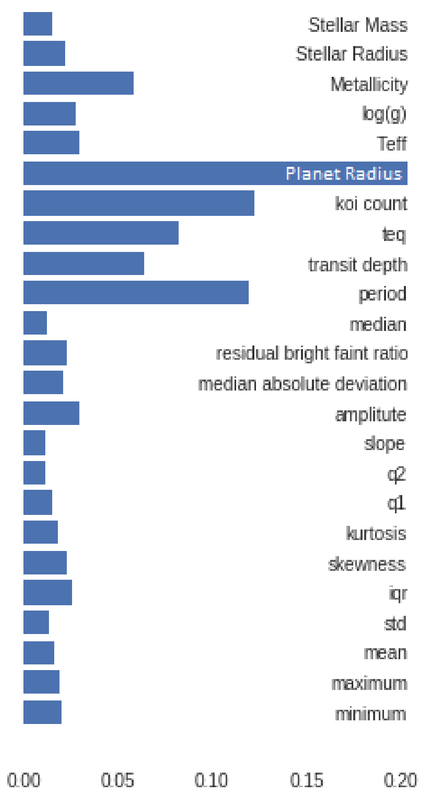
Para las técnicas automáticas de extracción de características se utilizaron dimensiones fijas: **5**, 10, 15, 20, 25, 50 para ICA, y **5**, 10, 25, 55, 100, 255 para PCA. El aumentar las características reflejó un mayor error en el conjunto de pruebas, posiblemente por sobre-ajuste.

	Modelos de Aprendizaje			
	k-NN	*Regresión Logistic*	*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
Metadata planeta	0.825	0.848	0.848	**0.870**
Metadata estrella	0.766	0.718	0.751	**0.766**
OwnFATS + metadata planeta & estrella	0.844	0.864	0.876	**0.883**

Figure 1: F1-score sobre las diferentes representaciones y los diferentes algoritmos de aprendizaje

Respecto al pre-proceso de los datos, el completar datos faltantes con ceros sobrepasó (~ 0.1 en F1) al ajuste lineal. Igualmente, se muestreó cada 3 valores en pos de reducir la sensibilidad de la curva y valores faltantes, desafortunadamente ésto empeoró los resultados. Adicionalmente, dado el desbalance de clases, se sub-muestreó la clase mayoritaria. Sin embargo, el mejor resultado (una mejora de ~ 0.1 en F1) fue el de poner peso en la función objetivo a las distintas clases.

5. Detalles



La importancia de los atributos del modelo mejor entrenado, **Random Forest**, en la mejor representación generada, OwnFats + metadata, se muestra en la imagen.
Se aprecia que los atributos más relevantes están asociados a los metadata del planeta y los menos relevantes a los estadísticos extraídos de la curva de luz.

6. Conclusiones

Se introdujo una nueva forma de decidir si un objeto en estudio (KOI) es realmente un exoplaneta utilizando métodos de Machine Learning sobre datos brutos. Los resultados se podrían mejorar en base a la elección de cuáles metadata utilizar.
Trabajo futuro Versión suave de la curva de luz (Mandel-Agol fit) y refinamiento de métodos de extrapolación y estimación de datos faltantes.

7. Resultados Finales

En pos de imitar el trabajo de los expertos, se realizó el etiquetado de los 1791 objetos candidatos que siguen en estudio por el staff de NexScI a Septiembre de 2017 obteniendo 975 Confirmados y 434 Falsos Positivos.
Se muestra el detalle del proceso de etiquetación:

- Mejor representación: OwnFATS + metadata planeta & estrella
- Modelo escogido para clase Confirmado: Random Forest ($depth = 15, T = 15$)
- Modelo escogido para clase Falso Positivo: SVM RBF ($C = 100$)

ID - KOI	**Etiqueta**	**Estrella (Conf.)**
K00601.02	*Falso Positivo*	Kepler 619 (2/3)
K00750.02	*No Clasificado*	Kepler 662 (1/3)
K01082.01	*Confirmado*	Kepler 763 (1/4)
K01082.02	*Falso Positivo*	
K01082.04	*Confirmado*	
K01236.04	*Confirmado*	Kepler 279 (2/3)
K01358.01	*Confirmado*	- (0/4)
K01358.02	*Confirmado*	
K01358.03	*Confirmado*	
K01358.04	*Confirmado*	
K01750.02	*Confirmado*	Kepler 948 (1/2)
K02064.01	*No Clasificado*	- (0/1)
K02420.02	*Confirmado*	Kepler 1231 (1/2)
K02578.01	*Falso Positivo*	- (0/1)
K02828.02	*Falso Positivo*	Kepler 1259 (1/2)
K03444.03	*No Clasificado*	- (0/4)
K07279.01	*Confirmado*	- (0/1)
K07378.01	*Confirmado*	- (0/2)
K07378.02	*Confirmado*	

Más en github.com/FMena14/ExoplanetDetection