Clasificación de alertas para el sistema Broker ALeRCE: El Clasificador de Curvas de Luz

Glosario

ALeRCE: Automatic Learning for the Rapid Classification of Events

ZTF: Zwicky Transient Facility. Es un proyecto de búsqueda sistemática de fenómenos astronómicos transitorios de duración corta (segundos a años) en el hemisferio norte. Abarca novas, supernovas, asteroides, cometas (por delante de estrellas). Es el sucesor del proyecto Palomar Transient Factory (PTF).

AIIWISE: Un atlas de astronomía, al parecer.

Balanced Random Forest: algoritmo de aprendizaje basado en decisiones sobre un árbol binario balanceado.

Estocástico: [proceso] que está sometido al azar y que es objeto de análisis estadístico.

LLST: una cámara brígida para un observatorio brígido

Sección 1: Introducción

Las variaciones de brillo de objetos astrofísicos ofrecen una idea clave de sus mecanismos físicos de emisión y de los fenómenos relacionados. En las estrellas, una pulsación - ya sea radial o no radial - puede resultar de un motor termodinámico operando en sus capas de ionización parcial, cuando las estrellas dentro de uno de los muchos llamados "viajes de inestabilidad" que se encuentran en el diagrama de Hertzsprung-Russell. Los eventos eruptivos pueden ser generados por material desprendiéndose de una estrella, o ocasionalmente agregándose a ellas, como es típico de protoestrellas y de objetos estelares jóvenes (YSO, Young Stellar Objects).

Eventos explosivos pueden ocurrir cuando el material se agrega en objetos más compactos, como los *enanos blancos* en el caso de variables cataclísmicas (CVs) o estrellas de neutrones en el caso de binarios de rayos X, o fusionadores de estrellas. Cambios en el brillo también pueden encontrar su origen en la rotación de estrellas, causada por características superficiales tales como las manchas estelares, y/o por formas de estrellas elipsoidales. Finalmente, los eclipses pueden ocurrir, dependiendo de la línea de visión del observador, debido a la presencia de compañeros binarios, planetas, y/o otros materiales circunestelares. Esas y otra clases de variaciones estelares son revisadas y sumarizadas, por ejemplo, en Catelan & Smith (2015), donde se pueden encontrar muchas referencias adicionales. En adición, hay una gran formación de transientes como lo son las kilonovas (), las supernovas (), y eventos de disrupción de marea (), que son balizas de episodios destructivos en la vida de una estrella. Las galaxias, por otro lado, también pueden estar presentes en una gran cantidad de fenómenos de variabilidad. En aquellos que involucran agujeros negros de agregación fuerte, por ejemplo, las variaciones se desarrollan debido a la naturaleza estocástica del disco de agregación, de la corona, y de la emisión de chorro, potencialmente relacionada tanto a las propiedades del agujero negro como a la estructura del material en la vecindad inmediata.

Para estudiar la variabilidad de objetos individuales en detalle y usar esta información para probar diferentes modelos físicos, las observaciones sobre un amplio rango de escalas de tiempo son esenciales, lo cual motiva a largas e intensivas campañas con una gran cantidad de objetivos (??). En los años recientes, encuestas cubriendo una parte importante del cielo, revisitando la misma región en escalas de tiempos de días a años, y conteniendo un gran número de objetos fortuitos, están disponibles como predecesoras de la Encuesta de Legado del Espacio y Tiempo del Observatorio Vera C. Rubin (Vera C. Rubin Observatory Legacy Survey of Space and Time).

Entre estas se encuentra la facilidad transiente de Zwicky (ZTF), que tuvo su primera luz en el 2017 y empleó una poderosa cámara con un campo de visión de 47 deg2 montada en el telescopio Samuel Oschin, de 48 pulgadas. ZTF está diseñado para fotografiar el cielo norte en su totalidad cada tres noches y escanear el plano de la Vía Láctea veinte veces por noche, a una magnitud limite de 20.5 en *gri*, dando paso a una amplia variedad de nuevos estudios de serie multibanda, en preparación para el LSST.

LLST, que está planeado para tener su primera luz en 2022, revolucionará la astronomía del tiempo, permitiendo por primera vez el estudio del transiente y de objetos variables sobre largos periodos de tiempo (del orden de 10 años) con más de 10000 visitas, llegando a magnitudes bajísimas (r en orden 24.5 para imagenes individuales del cielo completo cada 3 días, 26.1 para stacks anuales, y 27.5 en máxima profundidad), sobre una gran área del cielo (sobre 18000 deg2).

Dado el gran número de fuentes que ZTF y LSST observarán (entre 1 y 40 billones de objetos), es crítico desarrollar técnicas confiables y eficientes de variabilidad. Esta nueva información permitirá ver a través de degeneraciones que podrían existir a partir de la caracterización del color únicamente. Esas técnicas de selección deberían, idealmente, tomar ventaja de las curvas de luz multibanda provistas por encuestas como LSST y ZTF, y separar diferentes subclases de objetos variables y transientes sin la necesidad de un espectro óptico, lo cual sigue siendo muy caro de obtener para ejemplos grandes.

Esta nueva generación de telescopios de encuesta ampliada ha demostrado una creciente necesidad de sistemas de alerta astronómica sofisticados (es decir, sistemas que puedan detectar cambios en el cielo de un origen astrofísico). Esos sistemas involucran el procesamiento en tiempo real de data para la generación de alerta, la anotación y la clasificación de las mismas (hasta 40 millones de alertas por noche) en tiempo real y la reacción en tiempo real a alertas interesantes usando los recursos astronómicos disponibles (por ejemplo, vía Target Observation Managers, o TOMs). Para poder usar esos recursos de forma inteligente y eficiente, la comunidad astronómica ha estado desarrollando una nueva generación de sistemas de filtrado de alertas conocidos como "brokers". Un ejemplo de Broker generado por la comunidad es el proyecto ALeRCE (Automatic Learning for the Rapid Classification of Events). ALeRCE es una iniciativa guiada por un equipo interdisciplinario e interinstitucional de científicos de distintas instituciones, tanto en Chile como en los Estados Unidos. El principal objetivo de Alerce es facilitar el estudio de variables sin movimiento y objetos transientes.

ALerCE está actualmente procesando el sistema de alertas de ZTF, proveyendo de clasificaciones distintas variables y objetos transientes, en preparación para la era LSST. Dos modelos de clasificación están actualmente disponibles en la pipeline de ALeRCE: un **clasificador de *stamp* o clasificador temprano**, que usa una Convolutional Network en la primera detección de stamp de una fuente para clasificarla entre **cinco clases generales: variable star, active galactic nuclei, SN, asteroid, bogus,** y un **clasificador de curva de luz o *clasificador tardío***, que usa una variabilidad de características computadas a partir de las curvas de luz para clasificar cada fuente en subclases más específicas (actualmente 15) entre tres de las 5 clases principales.

En este trabajo se presenta la primera versión del clasificador de curvas de luz ALeRCE. Este clasificador usa varias características novedosas y emplea algoritmos de aprendizaje de máquinas que pueden tratar con el alto **desbalance de clases** presentes en los datos, siguiendo un esquema de dos niveles. Una meta clave de ALeRCE es proveer clasificación rápida de los transientes y objetos variables en un framework altamente escalable, por lo cual solamente se incluyen en este modelo características que pueden ser computadas rápidamente, evitando aquellas características que requieren más de un segundo para computar, según la infraestructura computacional actualmente a disposición ()(figura). La principal ventaja de este clasificador es que puede separar múltiples clases de transientes y objetos variables usando características computadas de datos reales, que podrían ser obtenidos a partir de los datos del LSST. Particularmente, el clasificador de curvas de luz puede tratar con múltiples clases de objetos variables estocásticos (incluyendo núcleo, host, y jet-domained active galactic niclei, YSOs, and CVs), which have been normally not included by previous classifiers that use real data and classify periodic and transient objects.

Este trabajo busca separar un gran número de clases (15) sin precedentes, tanto de transientes como de objetos variables usando data real (en lugar de usar solo data simulada). Los trabajos previos usando data real se ha nconcentrado principalmente en seleccionar ya sea una variedad de clases de estrella variable, diferentes clases de objetos variables, incluyendo estrellas variables y active galactic nuclei, o diferentes clases de transientes () () ().

Bajo nuestro mejor conocimiento, los tres trabajos previos han usado datos reales para clasificar transientes y objetos variables, aunque considerando un bajo número de clases: uno () usó data de encuestas HiT s para clasificar 8 transientes, active galactic nuclei y clases de estrellas variables, otro () usó data del OGLE y del OSC para clasificar 7 tipos de transientes y estrellas variables, otro usó CRTS para clasificar 6 transientes y objetos variables. Otros trabajos... blablabla.

En adición, este trabajo es el primer intento de separar tres clases distintas de active galactic nuclei (core dominated o quasi-stellar objects - QSO, host-dominated - AGN, jet-dominated - Blazar). Los trabajos previos se han concentrado principalmente en separar active galactic nuclei del resto.

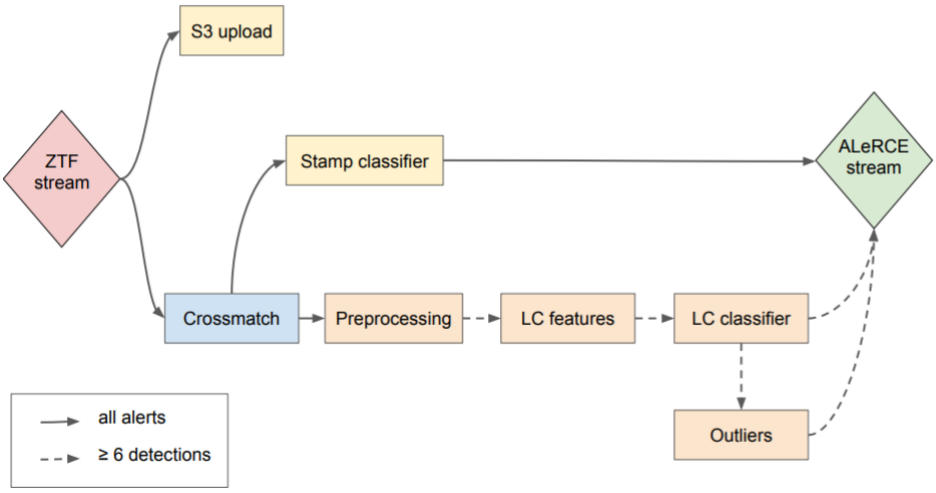
El paper se organiza de la siguiente manera:

- En la sección 2, se describe la data usada para el trabajo, el procedimiento para la construcción de una curva de luz, la taxonomía y el set etiquetado usado para entrenar el clasificador
- En la sección 3, se define el conjunto de características usadas por el clasificador de curvaz de luz.
- En la sección 4, se describen los distintos algoritmos de ML testeados para el clasificador.
- En la sección 5, se compara el rendimiento de los distintos modelos, y se reportan los resultados obtenidos para los conjuntos etiquetados y no etiquetados de ZTD.
- Finalmente, en la sección 6 se sumariza el paper, dando conclusiones y discutiendo los desafíos encontrados durante el desarrollo del clasificador y el trabajo futuro.

Sección 2: Datos

ALeRCE ha estado preprocesando el stream de ZTF (público) desde mayo del 2019, incluyendo fotometría *g* y *r*. La pipeline de ALeRCE se describe en detalle en el otro paper. A continuación se presenta una descripción breve del proceso de construcción de una curva de luz.

ALeRCE procesa los archivos de Avro Alert del ZTF. Esos archivos contienen metadata e información contextual para un evento singular, que se definen como flux-transient, reocurring flux-variable, o moving object. Para construir las curvas de luz, ALeRCE usa la fotometría de la difference-image y la reference-image (detecciones), posibles no-detecciones asociadas con el objetivo durante los últimos 30 días del evento, el puntaje real-bogus reportado por ZTF (rb, que va de 0 a 1, con valores más cercanos a 1 si la detección es confiable), y la clasificación morfológica del objeto más cercano obtenido por PanSTARRS1. En la Figura 1 se presenta una visión general del pipeline:



Los distintos niveles son:

1. Inyección: el stream de ZTF se inyecta usando Kafka.
2. S3 upload: los paquetes de alerta Avro se guardan en AWS S3 para acceso posterior.
3. Crossmatch: la posición de la alerta se usa para consultar catálogos externos.
4. Stamp classifier: las alertas de objetos nuevos se clasifican usando los cos *image cutout (stamps).*
5. Preprocessing: la fotometría asociada a una alerta dada se corrige para tomar en cuenta el uso de diferentes flujos de imagen, y se computan estadísticos simples con la curva de luz agregada.
6. Light curve features: se generan estadísticos avanzados sobre la curva de luz, cuando hay **al menos 6 detecciones en una banda dada.**
7. Light curve classifier: se aplica el clasificador de curvas de luz descrito en este trabajo.
8. Outliers: se aplica un algoritmo de detección de algoritmos.
9. ALeRCE Stream: las curvas de luz agregadas, anotadas y clasificadas se reportan al stream de Kafka.

En el paso 3 se utilizan varios catálogos, pero en este trabajo se usa principalmente el AIIWISE Public Source Catalog, invocando un match radius de 2 arcosegundos para obtener fotometría W1, W2 y W3.

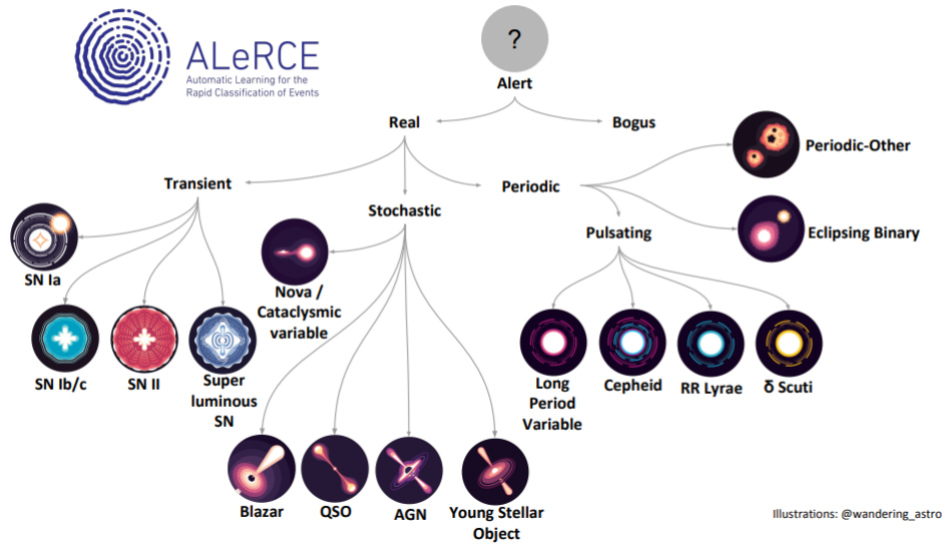


Figure 2. Taxonomy tree used in the current version of the ALeRCE light curve classifier.

The preprocessing procedure...

Taxonomía de clasificación

La primera versión de ALeRCE considera 15 subclases de variable y objetos transientes, presentados como un árbol de taxonomía definido en la Figura 2. La taxonomía se subdivide jerárquicamente de acuerdo a las propiedades físicas de cada clase y a las propiedades de variabilidad empírica de las curvas de luz, según sigue:

- Transient:
  - Supernova tipo Ia (SNIa)
  - Supernova tipo Ibc (SNIbc)
  - Supernova tipo II (SNII)
  - Supernova super luminosa (SLSN)
- Stochastic:
  - Galaxia Seyfert tipo I (AGN)
  - Quasar tipo 1 (QSO)
  - Blazar (Blazar)
  - Young Stellar Object (YSO)
  - Cataclysmic Variable/Nova (CV/Nova)
- Periodic:
  - Long-Period Variable (LPV)
  - RR Lyrae (RRL)
  - Cepheid (CEP)
  - Eclipsing binary (E)
  - Sigma Scuti (DSCT)
  - Other periodic variable stars (Periodic-Other)

La Figura 3 muestra ejemplos de curvas de luz de las diferentes clases consideradas por el clasificador, obtenidos usando data de ZTF.

Es importante mencionar que existe una cantidad menor de clases comunes que aún no han sido separadas en el árbol de taxonomías de ALeRCE, debido a que número de objetos cross-matcheados en esas clases es muy bajo para entrenar un buen modelo de clasificación (SNe IIb, TDEs, KNe, entre otros). Para los casos que caen en clases periódicas existe el catch-all "Periodic-

Other", mas para los casos transientes o estocásticos no existe tal cosa. Por el momento, esas clases perdidas se están agrupando en una o más de las clases ya existentes.

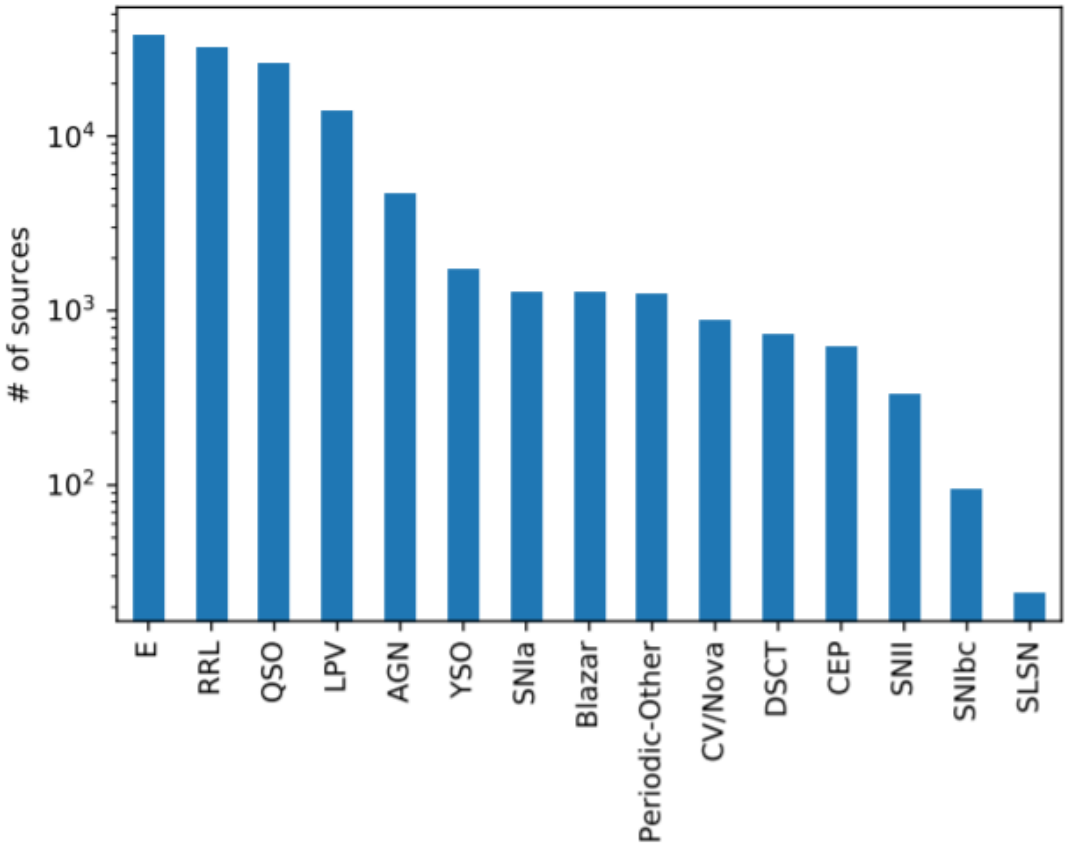


Figure 4. Number of sources per class for the labeled set, as reported in Table 1.

Table 1. Labeled set definition

Hierarchical Class	Class	# of sources†	Source Catalogs
Transient	SNIa	1272 (74.0%)	TNS
	SNIbc	94 (5.5%)	TNS
	SNII	328 (19.1%)	TNS
	SLSN	24 (1.4%)	TNS
	Total	1718	
Stochastic	QSO	26168 (75.4%)	MILLIQUAS (sources with class "Q")
	AGN	4667 (13.4%)	Oh2015, MILLIQUAS (sources with class "A")
	Blazar	1267 (3.6%)	ROMABZCAT, MILLIQUAS (sources with class "B")
	YSO	1740 (5.0%)	SIMBAD
	CV/Nova	871 (2.5%)	TNS, ASASSN, JAbril
	Total	34713	
Periodic	LPV	14076 (16.2%)	CRTS, ASASSN, *Gaia*DR2
	E	37901 (43.5%)	CRTS, ASASSN, LINEAR
	DSCT	732 (0.8%)	CRTS, ASASSN, LINEAR, *Gaia*DR2
	RRL	32482 (37.3%)	CRTS, ASASSN, LINEAR, *Gaia*DR2
	CEP	618 (0.7%)	CRTS, ASASSN
	Periodic-Other	1256 (1.4%)	CRTS, LINEAR
	Total	87065	

† Values in parentheses correspond to the fraction of sources of a given class (second column) within its corresponding hierarchical class (first column).

LOS DATOS ESTAN TERRIBLEMENTE DESNIVELADOS

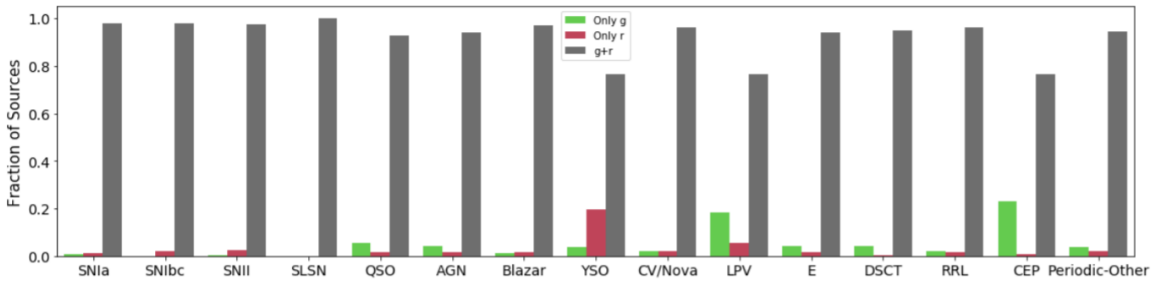


Figure 5. For the sources in the labeled set, this figure shows the fraction of sources in each class with photometry: only in the g band (green); only in the r band (red); or in both bands (grey). The reasons for the non-uniformity of coverage may be physical (strongly red or blue source) or organizational (survey focused on one band only). For most classes, the vast majority of the sources ($\gtrsim$92%) have photometric detections in both g and r; the exceptions are the YSO, LPV, and CEP classes, where only 76% of the sources have photometry in both bands.

Y NO PARA TODOS SE TIENEN LAS 2 BANDAS

Sección 3: Características usadas por el Clasificador

El clasificador de curvas de luz usa un total de 152 features. Se evita incluir aquellas que requieren de un largo tiempo de procesamiento, como lo son aquellas que requieren el uso de técnicas de Cadena de Markov Monte Carlo, puesto que una de las metas del clasificador es ser rápido y escalable. 142 de esas características se computan usando exclusivamente las bandas *b* y *r* de la data pública de ZTF.

Excluimos la magnitud media como una característica para evitar que cualquier bias en la distribución de magnitud del set etiquetado afecte la clasificación de fuentes que son más débiles o más brillantes. Las features obtenidas usando las magnitudes observadas por ZTF se denominan **detection features** (56 features en la banda g, 56 en la banda r y 12 features multibanda, con un total de 124 features), mientras que las features obtenidas computando la no-detección ZTF de magnitud 5sigma limits diffmaglims se denominan **non-detection features** (9 para cada banda g y r, dando un total de 18).

También se incluyeron como características las coordenadas galácticas de cada objetivo (gal_b y gal_l), los colores W1-W2 y W2-W3 de AIIWISE, y los colores g-W2, g-W3, r-W2, r-W3, donde g y r se computan como la magnitud media de la banda g y r para una fuente dada. Además, usamos información incluida en la metadata de los archivos Avro: el parámetro sgscore1, que corresponde a un puntaje de la fuente más cercana de PanSTARRS1 (estrellas morfológicas o galaxias). Dicho parámetro varía entre 0 y 1, donde un valor más cercano a 1 implica una *higher likelihood* (probabilidad más alta) de que la fuente sea una estrella. También se usa el parámetro rb, que corresponde a la mediana de real-bogus. Con esas 10 características extras, el número total de características usadas por el clasificador asciende a 152.

En este paper solo se consideran curvas de luz con 6 o más épocas en g o r. Si una fuente dada cumple con este criterio pero solo en una banda, se incluye en el análisis y las características asociadas a la banda para la cual no se tiene información se consideran como **valores -999.**

Detection Features

La mayoría de las features usadas por el clasificador de curvas de luz se computan usando las magnitudes obervadas en las bandas g y r (las *detecciones*). Hay 56 features computadas para cada banda y 12 usando una combinación de ambas, entregando un total de 124 features. La definición de todas esas features se encuentra en la Tabla 2, que se divide en tres bloques:

- El primer bloque contiene nuevas características incorporadas en este paper (novel features). Algunas de esas características se describen en la sección 3.1.1.

- El segundo bloque contiene features que corresponden a nuevas variantes de descriptores incluidos en otros trabajos. Algunos de ellos se describen en el Apéndice A.
- Finalmente, el tercer bloque incluye 22 features que vienen del paquete de Python Feature Analysis for Time Series (FATS).

Las features que terminan en _1 se computan usando la banda g, mientras que las features que terminan en _2 se computan utilizando la banda r, siguiendo la notación usada por los archivos Avro de ZTF.

Explicación de cada feature importante en el paper

Sección 4: Algoritmos de clasificación utilizados

El set etiquetado usado en este trabajo presenta un desbalance muy alto. Por ejemplo, los QSOs representan el 75,4% de las fuentes estocásticas, mientras que los casos de CV/Novae representan solo el 2.5%. Para sobrellevar esta situación, se buscaron algoritmos de Machine Learning disponibles en la literatura que estuviesen diseñados para mitigar el problema d desbalance. En particular, se trabajó con el paquete de Python de **imbalanced-learn**, que incluye implementaciones de una gran cantidad de algoritmos de re-sampling comúnmente usados para manipular datasets con un desbalance importante de clases. Los algoritmos disponibles en este paquete son totalmente compatibles con los métodos de **scikit-learn.**

El primer algoritmo a estudiar es Random Forest, algoritmo usado por el clasificador de curvas de luz. Para entrenar cada clasificador se dividió el dataset etiquetado en un set de entrenamiento del 80% y un set de testing del 20%, de **forma aleatoria estratificada,** preservando el porcentaje de samples de cada clase.

Balanced Random Forest

Un árbol de decisión es un algoritmo predictivo que usa una estructura de árbol para llevar a cabo particiones sucesivas de la data de acuerdo a un cierto criterio (por ejemplo, un valor de corte en uno de los descriptores o features) y produce posibles caminos de decisión, dando origen a una salida final para cada camino (las hojas del árbol). Los árboles de decisión comúnmente se usan para clasificar, aplicación en la cual cada hoja final se asocia con una clase dada. Los Random Forest (RFs en adelante) son algoritmos que construyen múltiples árboles de decisión, donde cada árbol se entrena usando un sub-sample aleatorio de elementos de un set de entrenamiento dado, seleccionando permitiendo la repetición **(bootstraping)** y usando selección aleatoria de características. La clasificación final se obtiene promediando las clasificaciones provistas por cada árbol. Este valor promedio puede ser interpretado como la probabilidad (Prf) de que el elemento de entrada pertenezca a la clase dada. Una de las principales ventajas de RF es que provee naturalmente un ranking de características para la clasificación, contando el número de veces que cada característica es seleccionada para particionar la data.

Chen (2004) propone un RF modificado que puede sobrellevar el problema de la data desbalanceada en clasificación. En este modelo cada árbol individual se entrena usando un subsample de el set de entrenamiento que se define generando un sample de bootstrap desde la clase minoritaria, y luego seleccionando aleatoriamente el mismo número de casos, con reemplazo, de las clases mayoritarias. El paquete **imbalanced-learn** implementa el clasificador RF balanceado propuesto por Chen. Para el clasificador de curvas de luz ALeRCE usamos el método **BalancedRandomForestClassifier**, seleccionando los hiper parámetros (número de árboles, máximo número de características por árbol, y máxima profundidad de cada árbol) con una **K-Fold Cross-Validation** (disponible en scikit learn), con k=5 folds y usando la macro-recall como métrica objetivo.

El acercamiento de clasificación en dos niveles

Considerando estructura jerárquica de la taxonomía, se decidió construir un clasificador RF balanceado con un esquema de dos niveles. El primer nivel consiste de un solo clasificador que separa las fuentes en tres clases principales. El segundo nivel consiste en tres clasificadores distintos, que resuelven cada clase del primer nivel en subclases. Luego usamos las probabilidades obtenidas de cada clasificador independiente para obtener la clasificación final.

En más detalle, el primer nivel (top level desde ahora) consiste de un único clasificador que clasifica cada fuente como periódica, estocástica, o transiente. El segundo nivel (bottom level desde ahora) consiste en tres clasificadores distintos: Transient, Stochastic, y Periodic. Las clases consideradas por cada una de esos tres clasificadores son las que se muestran en la Tabla 1 y Figura 2. Cada clasificador en el nivel más bajo se entrena usando un subconjunto de entrenamiento que tiene solamente las clases incluidas en la clase superior primaria (por ejemplo, el clasificador Transient solo incluye fuentes clasificadas como SNIa, SNIbc, SNII y SLSN). Es importante notar que esos cuatro clasificadores son independientes y procesan el mismo conjunto de features de entrada descritos en la sección 3. La clasificación final se construye multiplicando las probabilidades obtenidas para cada clase del nivel superior con las probabilidades individuales obtenidas por sus clasificadores correspondientes en el nivel inferior. Por ejemplo, las probabilidades del clasificador Transient se multiplican por P_top(transient), las probabilidades del clasificador Stochastic **(Ps)** se multilpican por P_top(stochastic), y las probabilidades del clasificador Periodic **(Ps)** se multilpican por P_pop(periodic) **{errata en el paper: usan la misma notación}** Denotamos el producto de esas probabilidades como P. Por ejemplo, la probabilidad de que una fuente dada sea un RRL corresponde al producto de su probabilidad de ser periodico (de acuerdo al nivel superior) y su probabilidad de ser un RRL (de acuerdo al clasificador Periodic, en el bajo nivel).

De esta forma, la suma de las probabilidades de las 15 clases para una fuente dada suma 1. Finalmente, la clase de un objeto dado se determina seleccionando la clase con la probabilidad P más alta. De aquí en adelante, nos referiremos a los resultados en el nivel inferior del clasificador como las predicciones finales.

El mejor rendimiento en cross-validation se obtuvo con los siguientes hiper parámetros:

- 500 árboles en cada clasificador
- máxima profundidad de árboles (los nodos se expanden hasta que todas las hojas sean puras)
- máximo número de features igual a la raíz cuadrada del total de features, excepto para el clasificador Stochastic, donde se usa el 20% de las features.

En la sección 5 se presentan los resultados obtenidos al aplicar el clasificador BRF a la data de ZTF.

Sección 5: Resultados

Sección 6: Sumario