LSST DESC Notes

The Photometric LSST Astronomical Time-series Classification Challenge (PLAsTiCC): Data set

Tarek Alam, Anita Bahmanyar, Rahul Biswas, Mi Dai, Lluis Galbany, Renée Hložek, Emille Ishida, Saurabh Jha, David Jones, Rick Kessler, Michelle Lochner, Ashish Mahabal, Kaisey Mandel, Juan Rafael Martinez Galarza, Alex Malz, Daniel Mutukrishna, Gautham Narayan, Tina Peters, Hiranya Peiris, and Kara Ponder

The Photometric LSST Astronomical Time Series Classification Challenge (PLAsTiCC) is an open data challenge to classify simulated astronomical time series data in preparation for the data from the Large Synoptic Survey Telescope (LSST), that will achieve first light in 2022. We briefly describe the PLAsTiCC data set that will be tested by the Kaggle team. This note will be updated for the full release of the data to the community.

1. Introduction

PLAsTiCC is a large data challenge where participants are asked to *classify astronomical time series data*. These simulated time series data, or 'light curves' are measurements of flux in different astronomical wavelength bands as a function of time for a large number of different astronomical sources, which make up different astronomical classes. The challenge is to classify each individual source as a member of such classes of objects. The time series data provided are simulations of what we expect from the upcoming LSST survey. For each object, the data provided includes summary information: its position on the sky, an estimate of its observed redshift (which correlates with its distance away from Earth), and other properties of the sky near the object. In addition, the light curve *photometry* data on the object is a table of fluxes at different times of observation, and at different wavebands (i.e. the average energy of the light within a range of wavelengths).

We go into more detail in the following section about the astronomical terminology used here.

In Figure **??**, we show three light curves for different types of objects.

The users are asked to classify the data into XX classes, XX-1 of which are represented in the training sample. The final class designation of ‘other’ is meant to capture objects that are hypothesized to exist but have never been observed and are thus not in the training set.

2. Astronomy Background

While we think of the night sky as static, it is filled with sources of light that vary in brightness on timescales from seconds and minutes to months and years.

Some of these events are classified as *transients*, and are the observational consequences of a large variety of astronomical phenomena. For example, the cataclysmic event that occurs when a supernova explodes generates a bright signal that fades with time, but does not repeat.

Other events are classified as *variables*, since they can vary their brightness in a periodic (or aperiodic) fashion, and originate from physical process governing high density regions of the Universe such as emission from the active galactic nuclei (AGN) at the hearts of galaxies, or as a result of geometric effects (e.g. eclipsing binary stars that alternately block out each others light from view).

These transient objects can provide important clues about themselves and their environment - as well as the evolution of the universe as a whole (e.g. type Ia supernovae provided the first evidence of the current accelerated expansion of the Universe which might be caused by dark energy).

Each different type of transient and variable provides a different clue that helps us study how stars evolve, the physics of stellar explosions, the chemical enrichment of the cosmos, and the accelerating expansion of the universe. Therefore, the proper classification

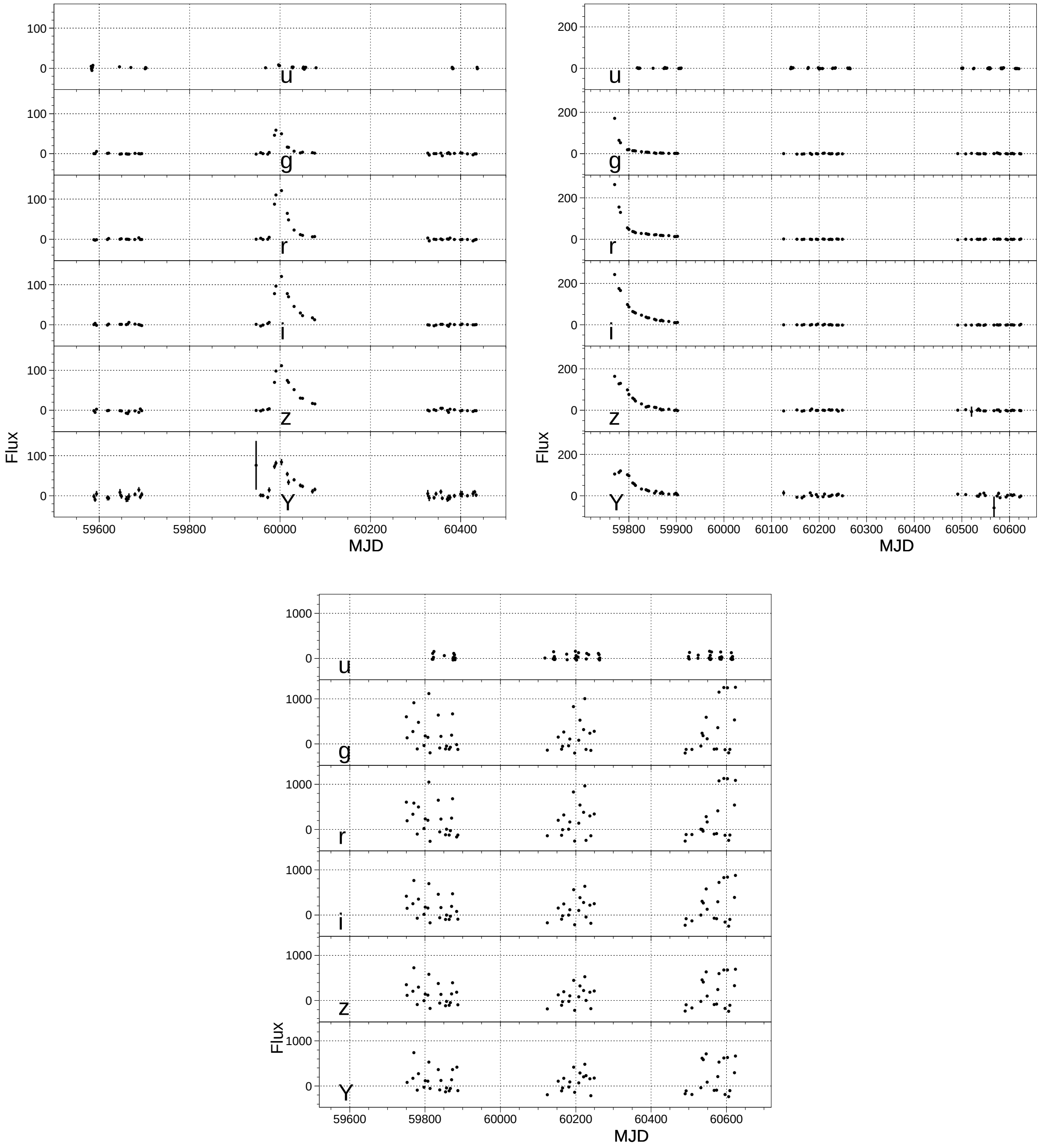

Figure 1. Example light curves in the PLAsTiCC data set. The three example objects display different changes in flux with time typical of real-world objects. They are either transient, and brighten suddenly before fading again into obscurity (top row) or they display flux variability, brightening and fading (bottom figure). This brightening can either be periodic or aperiodic. The top row also illustrates that the brightening of the flux can occur near the edges of the survey, and therefore may not include the full time period of brightening for the object. In addition, all three panels show that seasonal gaps and the instrument cadence of observations can introduce gaps in the light curve.

of transients is a crucial task in observational astronomy - specially in the light of large data volumes expected for the next generation of astronomical surveys.

The main aim of this challenge is: *can one classify astronomical transients and variables from a photometric light curve data set designed to mimic the data from the upcoming Large Synoptic Survey Telescope (LSST)?* Crucially, the classification will occur on a large test set, but the training data will be a small, and poorly representative training set, to mimic the challenges we face observationally.

2.1. Different ways of observing astronomical objects

Next we give more detail on LSST, and the challenge at hand. The two modes for characterising the light from the objects are called spectroscopy and photometry. Spectroscopy measures the flux per wavelength interval and is the modern equivalent of using a prism to separate a beam of light in its composite (e.g. rainbow) colours. It is a high resolution measurement which allows us to identify emission/absorption features indicative of specific chemical elements present in the object. Spectroscopy is also the primary tool that enables classification of astronomical transients and variables. Despite being paramount for the classification task, spectroscopy is an extremely time consuming process - with integration times ranging from 20 minutes to a few hours depending on the telescope and brightness of the source.

Given the volume of data expected from the upcoming large scale sky surveys, obtaining spectroscopy for every object is not sustainable. An alternative approach is to take an image of the object through different wavelength (band) filters, to determine the flux of the object. Classification is then performed on the light curves that result from those images.

Photometry records how bright the source is at a given moment. The photometric information is encoded as the flux (energy from the object). The photometric light curve has six pieces of information, namely the flux in six wavelength bands (named $ugrizY$) at any moment in time.

These photometric wavelength band fluxes are the integrals of the spectrum over the filter bandpasses of atmosphere and of the instrument divided by the energy of photons in the central wavelength of the filter. A sequence of photometric observations made at different times is called a light curve. It measures how the energy of the source evolves with time and can also be used to characterize different types of astronomical transients. As a consequence, for each object we will have a number of light curves in each filter (or band). Wavelengths are measured in units of Angstrom ($\mathring{A}$), where $1\mathring{A} = 10^{-10}m.$ Each band corresponds to a 'color', with a width of around 1000 $\mathring{A}$, with the full set ranging from $3000\mathring{A}$ (blue light) to $9000\mathring{A}$ (near infrared light).

The observations are affected by wavelength-dependent sky noise (due to e.g., moonlight and other sources). High-resolution spectroscopy carries thousands of information bits, and hence the challenge is to use the highly compressed photometric information to perform classification.

Unlike spectroscopy, photometry measures light from a large wavelength range simultaneously, and therefore collects more photons during observation, making it possible to measure light from objects at greater distances than with spectroscopy. For an object of given brightness (luminosity), the flux received on earth decreases with the distance to the object as

$$F = \frac{L}{4\pi d_L^2}. \tag{1}$$

The final connection is model cosmic distances through an expanding universe cosmological model. In such a cosmology of the universe, the connection distances to objects comes through the redshift, z. Redshift is an empirical quantity that is defined by measuring the difference in the observed wavelength λ_o of a given feature (e.g. in the spectrum described above) compared to the emitted wavelength λ_e, or

$$z = \frac{\lambda_o - \lambda_e}{\lambda_e}. \tag{2}$$

Just like the Doppler affect that acts on sound waves, redshifting is the analog for light. Using a spectrum to determine the redshift of an object gives the most precise result (with the smallest error σ_z). However photometry of the galaxy that 'hosts' the object, can also be used to determine a redshift for an object, the so-called photometric redshift, with a larger uncertainty. Photometric redshifts can include so-called catastrophic failures, where the redshift of the object is misassigned. These errors are rare (roughly 2% of the total number of objects), however they can pose serious problems for classification of objects.

2.2. The Large Synoptic Survey Telescope (LSST)

LSST is an ambitious telescope project under construction in Chile, scheduled to begin observations in 2022. With its powerful camera and wide field of view, it will be able to scan the whole sky visible from Chile once every three days. LSST will produce an unprecedented number of light curves by comparing images from day to day and looking for new objects not seen previously, and measuring the flux in those images. Once these transients are detected, we rely on agreements with other telescopes in order to acquire a small number of spectroscopic observations.

We will describe the data in the following sections, and discuss the metrics used to classify objects in a separate note.

3. The data

The photometric lightcurve data consist of non-homogeneously sampled, non-periodic time series with correlated errors obtained in several wavelength filters. A hdf file over all objects will be provided. The following data are provided:

The `PLAsTICC` data is provided in the form of a HDF5 file, which has two kinds of information. The first is a table listing each astronomical source in the data indexed by a unique identifier 'objid' which is a string. Each row of the table lists the properties of the source. These are:

- `objid`: the Object ID, unique identifier, string
- `ra`: right ascension, sky coordinate: co-longitude, units are degrees
- `decl`: declination, sky coordinate: co-latitude, units are degrees
- `mwebv`: a property of the Milky Way along the line of sight to the astronomical source, and is thus a function of the sky coordinates of the source `ra, decl` and has an

impact on the observations of objects along the same line of sight as described in subsection **??** and determines a wavelength (or `passband`) dependent dimming and redenning of the source light due to the Mikly Way dust.

- `hostgal_specz`: the spectroscopic redshift of the source. This is an extremely accurate measure of redshift, and is not measured for the test sample. Hence, these values are null in the test data.
- `hostgal_photoz` : The photometric redshift of the host galaxy of the astronomical source. While this is meant to be a proxy for `hostgal_specz`, there can be large differences between the two and should be regarded as a far less accurate version of `hostgal_specz`.
- `hostgal_photoz_err` : The uncertainty on the `hostgal_photoz`
- `sntype` : The class of the astronomical source as a string. (currently this variable is called 'sntype' and will be changed in future)

The second piece of information about the transients is its brightness as a function of time in different `passband`s, ie. light curve data. This is contained in a second table where each row corresponds to an observation of the source at a particular time and passand. This table includes the following information

- `mjd`: the time in Modified Julian Date (MJD) of the observation with a unit of day.
- `passband` : The specific LSST `passband` 'u' or 'g' or 'r' or 'i' or 'z' or 'Y' in which it was viewed. This is a categorical variable of the type string.
- `flux`: the measured flux (brightness) in the `passband` of observation as listed in the `passband` column. This is a float.
- `fluxerr`: the uncertainty on the measurement of the `flux` listed above.
- `photflag`: ignore.

A few caveats about the light curve data are in order.

- **Negative Flux** Due to the way the brightness is estimated, the flux may turn out to be negative for dim sources, where the true flux is close to zero.

- **saturated observations** The light curves include ‘saturated’ observations of sources, where the source is too bright to obtain a precise measured value. In such cases, the `flux` is set to 0., and the `fluxerr` is set to 10,000,000. Such an observation may not yield a value of `flux`, but it indicates that the source was extremely bright rather than extremely dim at the time of observation.

It is possible that the distribution of properties (as found in the header table) could be different for different classes of astrophysical sources. For example, sources that are extremely dim, may only be found in our own galaxy the Milky Way, and thus their redshifts will be close to zero, and their locations are likely to be clustered around the parts of the sky where the Milky Way is densely populated. Sources that are somewhat brighter, but still too dim to be see from large cosmological distances may be found at low redshifts only. On the other hand, extremely bright sources which are visible for extremely large distances may tend to be found at higher redshifts due to larger physical volume at high redshifts. On the other hand, the signature of a particular class of astrophysical source is to be found in its light curve, which is a degraded and compressed version of its spectral evolution.

3.1. Obtaining the data and Scoring a classification

The header table is located in the sample HDF5 file at the path `header`, while the light curve is stored at the path ‘objid’ in the sample HDF5 file. As part of the challenge, we provide an example Jupyter notebook to read in the data from an HDF5 file, and a notebook to compute the metrics for the challenge.

3.2. Training data

The training data follow the description above and have the properties and light curves of a set of 5000 astronomical sources and are meant to represent the kind of brighter objects for which obtaining expensive spectroscopy might be possible. The test data, meanwhile are supposed to be all of the data for which no expensive spectroscopy was obtained. Therefore, the test data has ‘NULL’ entries for the two columns `hostgal_specz` and `sntype`. Moreover, because of what they **do not form a fair sample of the test data set.** For example, The training data will necessarily be comprised of nearby, low redshift, brighter samples while the test data will contain higher redshift, fainter and more distant objects. It is entirely possible that there will be sources in test data, that do not have any counterparts in the training set.

4. Challenge participation

`PLAsTICC` challenge entries are required to classify each of the sources in the header file of the test data set based on their properties and light curves. The classification must be done though the assignment of probabilities $P(\mathrm{class}|\mathrm{data_i}, \mathrm{training} - \mathrm{data}, \mathrm{knowledge})$, the probability that the $i\mathrm{th}$ source in the test data is a member of the class class, based on $rmdata_i$, the combination of properties and light curve data for the object, the training data set, and any outside knowledge the participant may have acquired elsewhere. To specify the entry for a single source, the participant must provide the probabilities of that source belonging to each of the mutually exclusive (non-overlapping) ‘ classes in the training set, and of not belonging to any of the classes in the training set and therefore denoted by the `others` class. Obviously, high values of $P(\mathrm{class}|\mathrm{data_i}, \mathrm{training} - \mathrm{data}, \mathrm{knowledge})$ for a particular class and i indicate that the participant believes that the ith source is likely to be of class class. As true probabilities, the quantities $P(\mathrm{class}|\mathrm{data_i}, \mathrm{training} - \mathrm{data}, \mathrm{knowledge})$ must satisfy the following criteria:

$$
\begin{aligned}
0.0 \leq P(\mathrm{class}|\mathrm{data_i}, \mathrm{training} - \mathrm{data}, \mathrm{knowledge}) \leq 1.0 \qquad & \forall \mathrm{class}, \mathrm{i} \\
\sum_{\mathrm{class}} P(\mathrm{class}|\mathrm{data_i}, \mathrm{training} - \mathrm{data}, \mathrm{knowledge}) = 1.0 \qquad & \forall \mathrm{i}
\end{aligned}
$$

For a `PLAsTICC` entry to be valid, it must have these probabilities for each class and astronomical source, ie. an entry cannot leave out probabilities on any source, or class. To win the challenge, the entry must be a valid entry and minimise the PLAsTiCC metric score (which is described in a separate note included in this challenge). For a participant using a classifier which decides that an object is of a particular class rather than provide probabilities, the participant has to use their own prescription to define probabilities. For example, a source classified as the first class may be given a probability of 1.0 and all other classes probabilities of 0.0, or the first class may be assigned 0.9 and all other classes may be given uniform probabilities so that the probabilities sum to unity.

For example, if the challenge was to classify a set of 3 observations into two classes of ‘star’ or ‘galaxy’ classes (and an ‘other’ class), the returned classification table would be $3x3$ matrix:

While some members have been shielded from information about model specifics, the PLAsTiCC team involved in validating the data will not be able to participate in the challenge directly, and will only publish classifications on the data once the challenge has completed.

Object ID	$P(star)$	$P(galaxy)$	$P(other)$
1	0.6	0.3	0.1
2	0.3	0.3	0.4
3	0.55	0.4	0.05

Table 1. An example classification table for a challenge to classify 3 objects into 3 classes

4.1. Acknowledgments

The PLAsTiCC data relies numerous members of the astronomical community to provide models of astronomical transients and variables. These will be outlined in a paper to be published once the challenge is complete. While we cannot thank them by name at this stage (as this would no doubt identify the models included in the challenge), we acknowledge their contributions anonymously at this stage. This work was supported by an LSST Corporation Enabling Science grant, and a Dark Energy Science Collaboration Workshop support grant.

The DESC acknowledges ongoing support from the Institut National de Physique Nucléaire et de Physique des Particules in France; the Science & Technology Facilities Council in the United Kingdom; and the Department of Energy, the National Science Foundation, and the LSST Corporation in the United States. DESC uses resources of the IN2P3 Computing Center (CC-IN2P3–Lyon/Villeurbanne - France) funded by the Centre National de la Recherche Scientifique; the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231; STFC DiRAC HPC Facilities, funded by UK BIS National E-infrastructure capital grants; and the UK particle physics grid, supported by the GridPP Collaboration. This work was performed in part under DOE Contract DE-AC02-76SF00515.