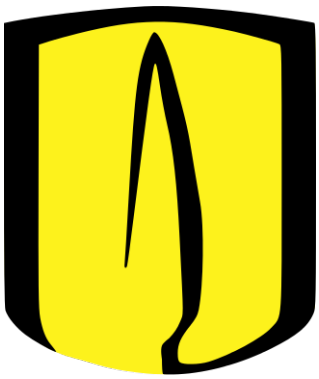

Universidad de los Andes

Undergraduate Thesis Proposal

Astronomical transient object classification using machine learning and deep learning techniques on real data

Author:
Mauricio Neira

Supervisor:
Marcela Hernández Hoyos, Ph.D.

February 11, 2019

Contents

1 Introduction

This project focuses on automatic transient object detection. This section will briefly outline the basic concepts needed to understand the motivation and reasoning behind it.

1.1 Astronomy

Astronomy is the field of science that studies celestial objects i.e. starts, planets, matter, etc. Outside of earth. These objects emit electromagnetic radiation all across the electromagnetic spectrum. Measurements of this radiation have been made with a wide variety of instruments and a lot of data has been recorded.

Usually, the data is recorded in *surveys* where various telescopes scan the sky through many days. Of of these surveys is the subject of this papers investigation known as the "Catalina Real Time Survey"[1]. In this database, the intensity of radiation of various objects was recorded at different time intervals. The group of points collected for a particular object is known as a light curve.

1.1.1 Light curves

After recording the intensity of emitted radiation of an object several times, it is possible to plot the result in a scatterplot. The following image is an example of a light curve of a type 1a supernova:

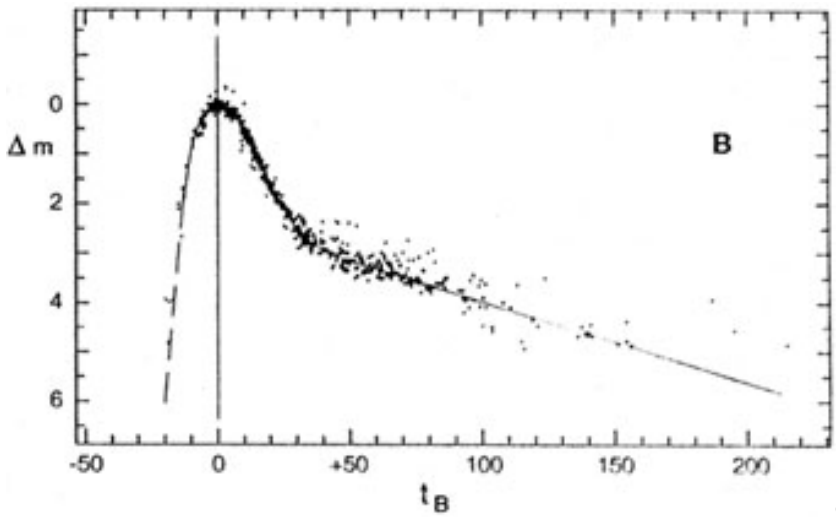

Figure 1: Light curve for a supernova type 1a [2].

Light curves hold a lot of information about the behavior of the object and can be used to identify the type of object [3, 4]. A big part of this project is dedicated to classifying objects based on their light curve.

1.1.2 Supernovae

Supernovae, specially type 1a supernovae, are of immense importance in present day astrophysics. They are used as "standard candles"[1] to estimate the rate of expansion of the universe [5]. So far, they have been used to pin down the cosmological constant (also known as Hubble's constant) and demonstrate the accelerating expansion of the universe. This has had a huge impact on our understanding of the universe. Particularly because it "impl[ies] the existence of a nearly uniform component of dark energy"[6] across space.

Finding and studying supernovae not only sheds light into their nature, it has implications that stretch out to the inner workings of the universe.

[1] A standard candle is a celestial object whose properties are optimal to measure large distances in the universe with high precision.

1.2 LSST

The Large Synoptic Survey Telescope is a project that will begin operating in 2020 and has four main goals[7]:

1. Probing dark energy and dark matter
2. Taking an inventory of the Solar System
3. Exploring the transient optical sky
4. Mapping the Milky Way

Unlike other telescopes of its nature, from the moment it begins operation, the LSST will release its data to the public. It plans on gathering 15 terabytes of data every night, generating alerts and transferring data with a 60 second delay [7].

With that amount of data, there simply aren't enough experts to classify all of the objects of interest. New automated techniques are needed to correctly classify all of the observed objects and notify the corresponding interested scientists.Consequently, the motive behind this project is to develop an automatic classifier of astronomical objects.

2 Problem description

As stated in the previous section, the motivation behind this project is to create an algorithm that automates the classification of astronomical objects. The classification can be done directly on the images released by the CSS [1] or on the light curves extracted from the images.

2.1 Classification of transient objects from light curves

One of the main aims to develop an algorithm that correctly maps the light curve of an object to its corresponding category. It is unrealistic to claim that the function will correctly classify every single light curve it is provided. Nevertheless, it does aim to minimize classification errors.

There will be 2 main approaches to achieve this goal. The first will be to transfer the light curves into a feature space that attempts to extract many geometric properties of the light curves. After that, several classifiers will be applied over the data points in the feature space. These include random forests and feed forward neural networks.

The second approach involves feeding the points to a recurrent neural network that will in turn be fed into a feed forward neural network.

2.2 Classification of transient objects from images

The second main aim of this proposed thesis is to develop an algorithm that correctly classifies astronomical objects from a series of *images*.

In brief, the aim is to implement a convolutional neural network that extracts descriptors from the images. Following this procedure, the extracted descriptors will be fed into a recurrent neural network that will have a fully connected layer attached at the end. The output of this architecture will be an n dimension vector corresponding to the n classes that the astronomical object can belong to. See section 5.3.2 for further details.

3 Project Background

3.1 Diego's thesis - 2018-10

In the first semester of 2018, Diego Alejandro Gómez Mosquera worked on "Astronomical transient event recognition with machine learning" using a variety of traditional machine learning classifiers on a feature space calculated from light curves [8].

The author focused on creating a feature space robust enough to distinguish the different transient classes. This was achieved primarily through geometric parameters that were extracted from the light curves. The features used and their importance during classification (as found on the author's thesis) for 8-class classification using a random forest classifier were:

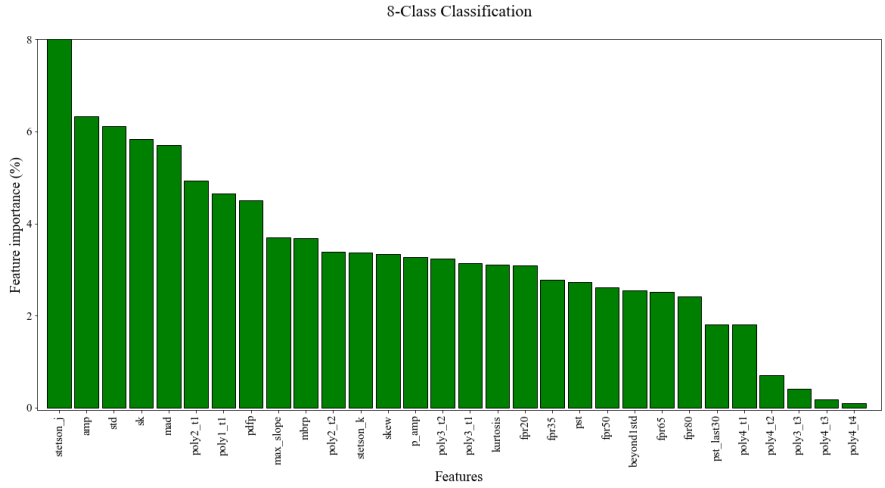

Figure 2: Feature importance when classifying light curves into 8 classes.

The classification metrics from the authors thesis will not be presented here as their validity is being revised for publication.

3.2 Research assistant work - 2018-20

During the second semester of 2018, I enrolled as a research assistant of Juan Pablo Reyes, Marcela Hernandez's Ph.D. student. During this time, several results relevant to this thesis were achieved.

3.2.1 Extension of Diego's work

Diego's classification used the only filter provided by CSS [1], the r filter. Naturally, one would like to extend the classification algorithm to include a variety of filters. The first extension of the algorithm was to include an arbitrary number of filters.

Since CSS's light curves only contain data on the r filter[1], a combination of data from the CFHT survey[9] and simulations run by Juan Pablo Reyes was used instead . The data had 4 filters in total : {r,i,g,z}.

It is worth noting that the original data for the algorithm was in magnitude while the CFHT data was in flux. Early experimentation demonstrated that classification in flux was very poor as it distorted the geometric properties of the light curves[2]. Additionally, considerable problems arose when converting flux to magnitude. Empirically, the classification metrics decreased. This is one of the main motivations for using the raw data from the light curves.

3.2.2 PLAsTiCC - Kaggle competition

In the second semester of 2018, the LSST team launched a competition based on simulated data[10]. The data included 6 filters $\{u, g, r, i, z, y\}$ and a total of 14 numbered but unlabeled classes. The multifilter approach from the previous section was implemented with a varying number of filters. Intuitively and experimentally, as demonstrated in the following table, the classification metrics improved as more filters were added:

		R	RG	RI	All filters	RNN -Red	Original quantity
Area under PR curve	**92**	0.94	0.95	0.95	0.97	0.60	239
	88	0.91	0.95	0.96	0.97	0.66	370
	42	0.35	0.34	0.34	0.38	0.25	1193
	90	0.57	0.57	0.6	0.72	0.52	2313
	65	0.76	0.7	0.75	0.83	0.54	981
	16	0.95	0.97	0.98	0.99	0.85	924
	67	0.05	0.1	0.13	0.21	0.04	208
	95	0.16	0.2	0.13	0.25	0.08	175
	62	0.1	0.13	0.15	0.28	0.12	484
	15	0.21	0.34	0.23	0.72	0.13	495
	52	0.07	0.05	0.07	0.06	0.05	183
	6	0.25	0.19	0.39	0.52	0.18	151
	64	0.13	0.15	0.07	0.19	0.05	102
	53	0.46	0.66	0.64	0.87	0.03	30

Figure 3: Classification results for the PLAsTiCC competition. The numbers in the columns correspond to the area under the precision recall curve for each one of the classes. Numbers close to red are shaded red while numbers close to 1 are not shaded.

Apart from the multifilter classification, a recurrent neural network was trained for a brief period of time. It is clear the metrics for this classifier do not compete with the multifilter approach but the neural network had at

[2]Conversion between these units involves a logarithmic scale thus the y-axis is greatly distorted, leading to inaccurate geometric fitting of data.

most 20 epochs to train. One of the objectives throughout this thesis will be to further train the RNN and improve its architecture for optimal results.

4 General objective

The general objective of this thesis is to develop an algorithm that can classify, as best as possible, astronomical objects into their corresponding class. Based on the catalina sky survey, the possible classes are:

- Active galactic nuclei
- Blazar
- Cataclysmic variable
- Flare
- High proper motion star
- Supernova
- Other (all transient objects not listed in the categories above)
- Non-transient

Two approaches will be taken to manage this objective. The first will be to use the light curve information as input. The second will be to use the whole images as the input.

5 Specific objectives

5.1 Improvement of the light curve feature space

5.1.1 Feature pruning

After seeing the results in 3.2.1, it was clear that the higher order coefficients in the polynomial fits did not contribute significantly to the correct classification in the feature space. In fact, these features could be harming the classification process. Thus, coefficients resulting from 3^{rd} and 4^{th} degree polynomial fitting will be removed and the random forest algorithm will

be rerun. The classification metrics should improve but experimentation is needed to confirm the hypothesis.

5.1.2 Addition of supernovae specific metrics

To improve the classification of supernovae, metrics that target their specific light curve behavior are needed. In particular, there are functions that are known to approximate the light curve produced by a supernova. These functions are presented below:

SALT2 is "an empirical model of Type Ia supernovae spectro-photometric evolution with time"[11]. This model should be better adjusted for type 1A supernovas than the rest of objects. The mathematical model of the function is the following[11]:

$$F(S, N, p, \lambda) = x_0 \times [M_0(p, \lambda) + x_1 M_1(p, \lambda) + ...] \times exp[cCL(\lambda)]$$

Skewed Gaussian fits of the form:

$$f^k(t) = A^k \frac{exp - (t - t_0^k/\tau_{fall}^k)}{1 + exp - (t - t_0^k)\tau_{rise}^k}$$

have been shown to resemble well a generalized supernova curve [12]. It should also approximate the shape of supernova light curves better than those that belong to other classes.

When these functions are fit to supernovae data, the resulting χ^2 should be considerably lower than the χ^2 calculated from non-supernovae fits. Consequently, the addition of the two χ^2 values from each of the functions should improve the binary classification of supernovae but as stated above, experimentation is needed for verification.

5.2 Deep learning on light curves

For the time being, the classification pipeline has had 3 steps:

1. Clean and filter light curves

2. Calculate features from clean light curves

3. Classify the objects on the feature space

When the feature space is calculated, a large quantity of information is lost. The features might be good descriptors of the objects but they will never be able to encapsulate the point-by-point data that the light curves have. Additionally, traditional machine learning methods, like the random forest classifier previously implemented, are unable to handle varying length input. To take advantage of all the information present in the light curves, a new approach is needed.

Recurrent neural networks (RNN) have been shown to be able to classify sequential data to a good extent. Speech recognition has been one of the fields with most progress[13]. A RNN like the long short term memory (LSTM) RNN or the gated recurrent unit (GRU) RNN are state of the art RNN's that seem promising for this purpose. Thus, several implementations of these networks will be carried out along with their hyperparameter tuning.

5.3 Deep learning on images

5.3.1 CNNs

So far, all the classification has been done on the objects' photometric light curves. The light curves, however, are not the raw data. These were calculated from the images that were taken by the telescopes. Specifically, the light curves werre calculated from the catalina real time survey images [14]. Ideally, to minimize data loss, the classification process should be done on the raw data i.e. the images.

The state of the art algorithms used for classifying images are known as convolutional neural networks (CNNs) [15] and have had wide success on large variety of problems.

The first step to correctly classify the images will be to implement a baseline CNN algorithm. Once its classification metrics are established, a thorough search through the hyperparameter space will be carried out to maximize the classification metrics.

5.3.2 CNNs and RNNs

Implementing a succesful CNN is only half of the problem. Each object has multiple images taken at different instances in time. To fully exploit the available information, multiple images need to be used as input for each object.

Since the amount of images per object is not fixed, a RNN will need to be used. Thus, the CNN will extract descriptors from the images and then those descriptors will be fed in chronological order along with the date of when the image was taken into the RNN for classification. The architecture is depicted below:

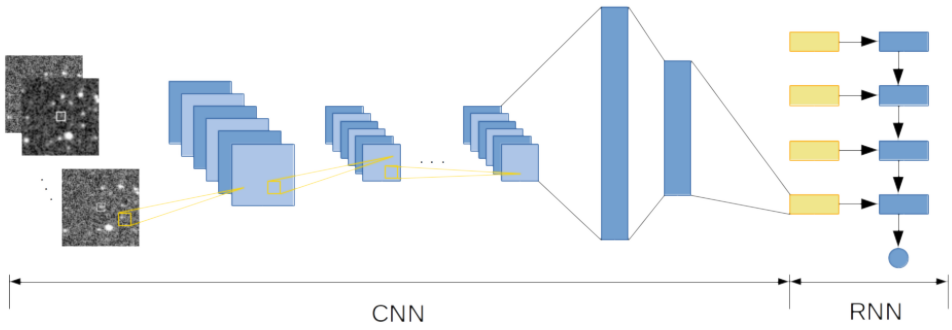

Figure 4: Architecture for classifying astronomical objects from multiple images.

6 Activities and schedule

Time	Activities
28 Jan - mid February	Improvement of feature space from light curves
mid February - March	1. Implementation and hyperparameter tuning of CNN for extraction of descriptors from images 2. Use of RNN to classify objects from light curves
March - April	Implement RNN on top of previously implemented CNN. Further tuning of previous neural networks.
April - May	Documentation

7 Expected results

With respect to the light curve based approach, several improvements are expected:

First of all, the addition of the χ^2 values for the SALT2 and skew gaussian fits should improve the classification of supernovas.

Secondly, if the RNN used to classify the light curves is let to learn for an increased number of epochs, an increase in the area under the PR curve for all of the classes is expected based on the results found in 3.2.2.

On the other hand, with regard to the classification of objects using the images as inputs, considerable improvements in classification metrics is expected. This expectation is based on the fact that the image as whole has a lot more information about the object than the light curve alone.

References

1. Drake, A. J. *et al.*
First Results from the Catalina Real-Time Transient Survey.
The Astrophysical Journal **696,** 870 (2009).

2. *Type Ia Supernovae as Standard Candles*
https://ned.ipac.caltech.edu/level5/Branch2/Branch2_1.html.

3. *Light Curves - Introduction*
https://imagine.gsfc.nasa.gov/science/toolbox/timing1.html.

4. *About Light Curves — Aavso.Org*
https://www.aavso.org/about-light-curves.

5. Goliath, M., Amanullah, R., Astier, P., Goobar, A. & Pain, R.
Supernovae and the Nature of the Dark Energy.
Astronomy & Astrophysics **380,** 6–18 (2001).

6. Perlmutter, S., Turner, M. S. & White, M. Constraining Dark Energy with Type Ia Supernovae and Large-Scale Structure.
Physical Review Letters **83,** 670 (1999).

7. Ivezic, Z. *et al.* LSST: From Science Drivers to Reference Design and Anticipated Data Products. *arXiv preprint arXiv:0805.2366* (2008).

8. Gomez, D. *Astronomical Transient Recognition Using Machine Learning and Catalina Real Time Survey Dataset.: Diegoalejogm/Crts-Transient-Recognition* Sept. 2018.

9. Canada, G. o.C.N.R. C. *National Science Infrastructure (NRC Herzberg, Programs in Astronomy and Astrophysics) - National Research Council Canada* eng.
http://www.cadc-ccda.hia-iha.nrc-cnrc.gc.ca/en/cfht/. Contact Information; Organizational Description; Promotional Material.
Apr. 2013.

10. *PLAsTiCC Astronomical Classification*
https://kaggle.com/c/PLAsTiCC-2018.

11. Guy, J. *et al.* SALT2: Using Distant Supernovae to Improve the Use of Type Ia Supernovae as Distance Indicators.
Astronomy & Astrophysics **466,** 11–21. ISSN: 0004-6361, 1432-0746 (Apr. 2007).

12. Möller, A. *et al.* Photometric Classification of Type Ia Supernovae in the SuperNova Legacy Survey with Supervised Learning. *Journal of Cosmology and Astroparticle Physics* **2016,** 008–008. ISSN: 1475-7516 (Dec. 2016).

13. Graves, A., Mohamed, A.-r. & Hinton, G.
Speech Recognition with Deep Recurrent Neural Networks
in *Acoustics, Speech and Signal Processing (Icassp), 2013 Ieee International Conference On* (IEEE, 2013), 6645–6649.

14. http://nesssi.cacr.caltech.edu/catalina/CRTSII_SNCV.html#table26.

15. Wang, C. & Xi, Y.
Convolutional Neural Network for Image Classification.
Johns Hopkins University Baltimore, MD **21218.**