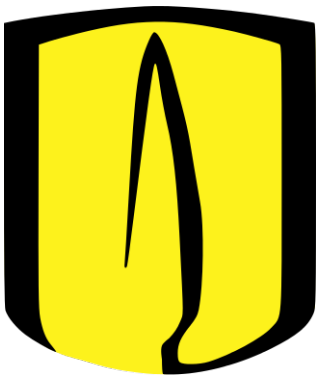

Universidad de los Andes

Undergraduate Thesis Proposal

Transient object classification using machine learning and deep learning techniques on real data

Author:
Mauricio Neira

Supervisor:
Marcela Hernández Hoyos, Ph.D.

February 5, 2019

Contents

1 Introduction

1.1 Transient objects

Two groups [1,2] have presented strong evidence that the expansion of the Universe is speeding up, rather than slowing down. It comes in the form of distance measurements to some fifty supernovae of type Ia (SNe Ia),

some stufff [**SN˙darkEnergy**]

1.1.1 Supernovae

1.2 LSST

2 Problem description

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 random forests on a feature space calculated from light curves [1].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 (as found on the author's thesis) were:

- skew: Skewness.
- kurtosis: Kurtosis.
- small kurtosis: Small sample kurtosis.
- std: Standard deviation.
- beyond1std: Percentage of magnitudes beyond one standard deviation from the weighted mean. Each weights is calculated as the inverse of the corresponding photometric error.

- stetson j: The Welch-Stetson J variability index [39]. A robust standard deviation.
- stetson k: The Welch-Stetson K variability index [39]. A robust kurtosis measure.
- max slope: Maximum absolute slope (delta magnitude over deltatime) between two consecutive observations.
- amplitude: Difference between maximum and minimum magnitudes.
- median absolute deviation:from the median magnitude.
- median buffer range percentage: Percentage of points within 10% of the median magnitude.
- pair slope trend: Percentage of all pairs of consecutive magnitude measurements that have positive slope.
- percent amplitude: Largest percentage difference between the absolute maximum magnitude and the median.
- percent difference flux percentile: Ratio of F 5,95 and the median flux.
- flux percentile ratio mid20: Ratio F 40,60 /F 5,95
- flux percentile ratio mid35: Ratio F 32.5,67.5 /F 5,95
- flux percentile ratio mid50: Ratio F 25,75 /F 5,95
- flux percentile ratio mid65: Ratio F 17.5,82.5 /F 5,95
- flux percentile ratio mid80: Ratio F 10,90 /F 5,95
- poly1 a: Coefficient of the linear term in monomial curve fitting.
- poly2 a: Coefficient of the cuadratic term in cuadratic curve fitting.
- poly2 b: Coefficient of the linear term in cuadratic curve fitting.
- poly3 a: Coefficient of the cubic term in cubic curve fitting.
- poly3 b: Coefficient of the cuadratic term in cubic curve fitting.

- poly3 c: Coefficient of the linear term in cubic curve fitting.
- poly4 a: Coefficient of the quartic term in quartic curve fitting.
- poly4 b: Coefficient of the cubic term in quartic curve fitting.
- poly4 c: Coefficient of the cuadratic term in quartic curve fitting.
- poly4 d: Coefficient of the linear term in quartic curve fitting.

The results from the authors thesis will not be presented here as a bug overrating the classification was found as will be discussed in the following section.

3.2 research internship?? No se como ponerle a esto 2018-2

3.2.1 Improvement on Diego's work

Diego's work was continued and improved during this semester with the mentorship of Marcela Hernández, Jaime Forero and Pablo Arbelaez.

3.2.2 PLAsTiCC - Kaggle competition

4 General objective

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, **REVISAR ESTO**, 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"[2]. This model should be better adjusted for type 1A supernovas than the rest of objects. The mathematical model of the function is the following[2]:

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

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 [3]. 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[4]. 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. **AGAIN, LINK HERE**. 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) [5] 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:

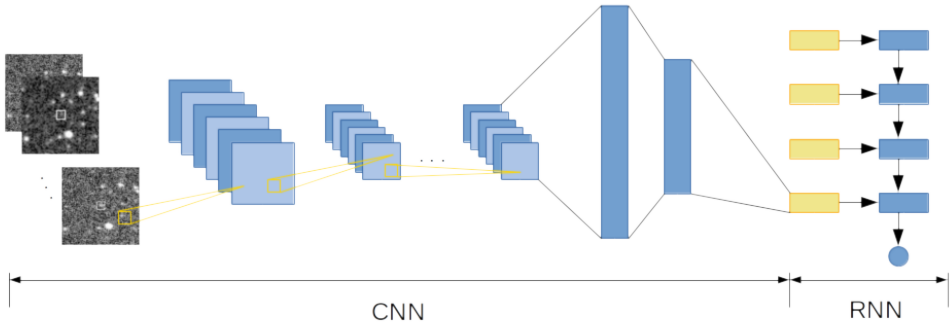

Figure 1: Architecture for classifying astronomical objects from multiple images.

DIAGRAM?

6 Activities and schedule

7 Expected results

References

1. Gomez, D. *Astronomical Transient Recognition Using Machine Learning and Catalina Real Time Survey Dataset.: Diegoalejogm/Crts-Transient-Recognition* Sept. 2018.
2. 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).
3. 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).
4. 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.
5. Wang, C. & Xi, Y. Convolutional Neural Network for Image Classification. *Johns Hopkins University Baltimore, MD* **21218.**