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.

January 31, 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 [**diegoThesis**].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 **??**, **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 **REFERENCES AND FUNCTION NEEDED HERE**

Skewed Gaussian **REFERENCES AND FUNCTION NEEDED HERE**

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 Machine learning on light curves

5.3 Deep learning on light curves

5.4 Deep learning on images

6 Activities and schedule

7 Expected results