Machine Learning Classification of Multi-band Supernovae Light Curves

Pablo Huijse[1,2], Pablo A. Estévez[1,2], Giuliano Pignata[1,3]

[1] *Millennium Institute of Astrophysics*
[2] *Departamento de Ingeniería Eléctrica, Universidad de Chile*
[3] *Departamento de Ciencias Físicas, Universidad Andres Bello*

Introduction

The precise classification of **Supernovae (SNe)** is critical for research in cosmology and astrophysics. Accurate classification of SNe can be achieved by studying their spectra, but spectroscopy is expensive to perform. On the other hand, photometric measurements are cheaper and available in much greater quantities.

Data Mining techniques based on **Machine Learning (ML)** methods have proved to be useful in a plethora of big-data astronomical problems [1]. At its simplest, supervised ML methods use a-priori knowledge (training dataset) to learn a model which is used to predict the class of unobserved data (test dataset).

In this work we develop a simple pipeline to process and classify multi-band light curves from the **Supernovae Photometric Classification Challenge (SNPCC)** [2] based on **Gaussian Processes (GP)** [3] and **Decision Trees (DT)** [4,5]. The results are compared to other methods found in the literature [6-10].

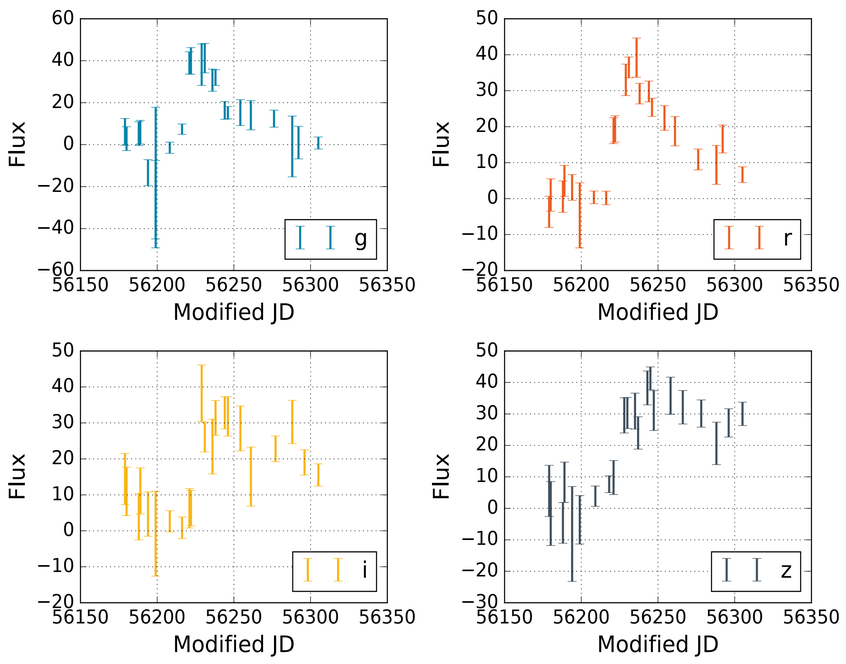

Fig 1. Multi-band (griz) light curve of SNPCC object DES_SN000017, it corresponds to a type II SN

Method

The SNPCC light curves were simulated following DES specification [2]. The objective of the competition was to classify 20,216 light curves (photometric subset) with a training set of 1,103 spectroscopically-confirmed light curves.

The following procedure is used to obtain a feature vector from a light curve

- **Fitting:** A GP with zero-mean, heteroscedastic errors, and squared exponential covariance is fitted to each band
- **Universal phasing:** The instant of maximum flux in the *r*-band is subtracted from the time vector
- **Normalization:** The predicted flux is divided by the maximum flux in the *r*-band
- **Sampling:** The mean posterior of the model is sampled in [-30, 90] days

The sampled posteriors are concatenated into a feature vector

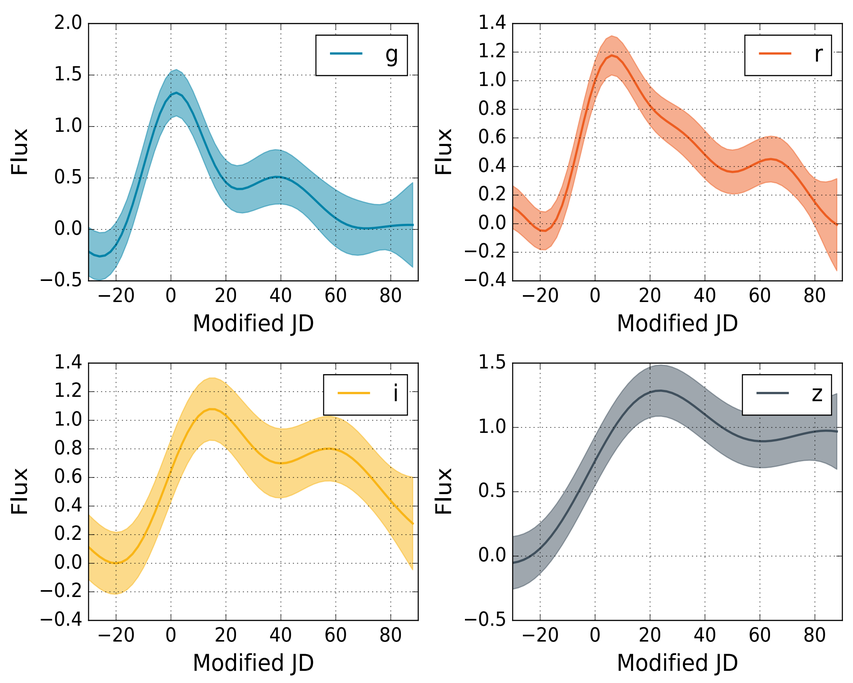

Fig 2. Result after fitting, phasing, normalizing and sampling for SN000017. The mean posteriors (solid line) are used as features

The features plus the photometric redshift (photo-z), are used to train a ensembles of DTs. We test **Random Forests (RF)** [4] and **Gradient Boosted Trees (GBT)** [5], which correspond to bagging and boosting based ensembles, respectively.

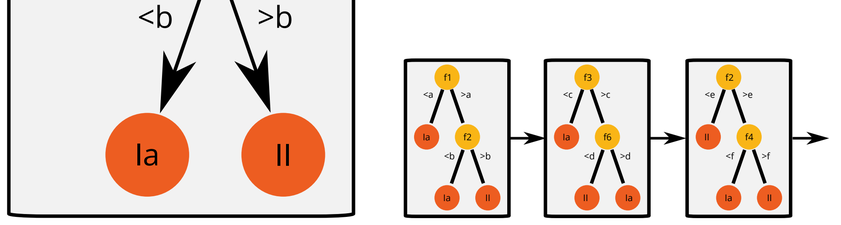

Fig 3. (a) Decision tree, (b) Bagging ensemble, (c) Boosting ensemble

Results

We train our method on randomly-sampled subsets of 1,103 light curves (5% of the total dataset) as others have done [6-10]. Results are measured in terms of precision and recall. Fig, 4 shows that the best performance is achieved by GBT.

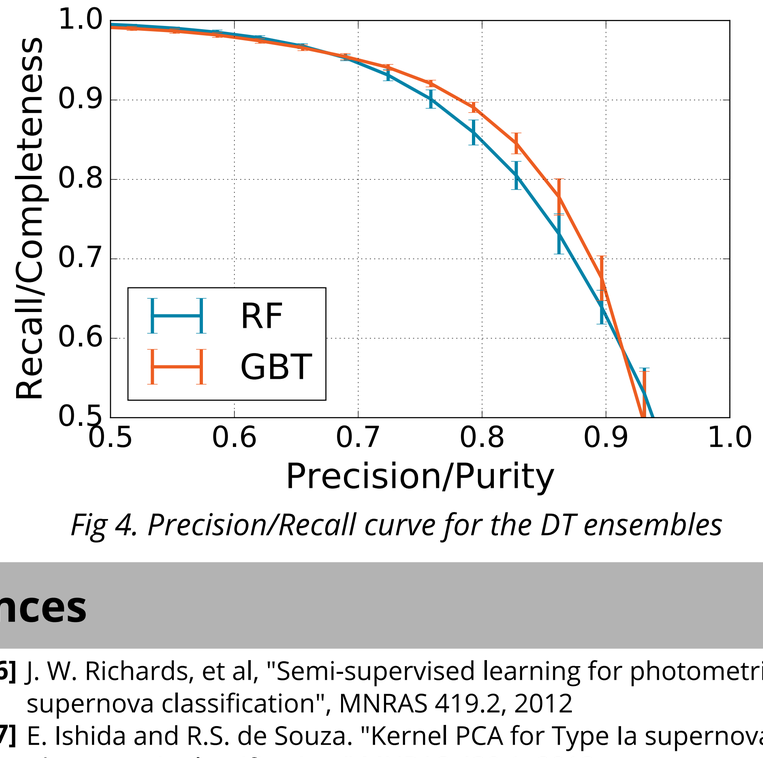

Fig 4. Precision/Recall curve for the DT ensembles

	Precision	Recall	F1-score
Karpenka et al. 2012	0.700	0.750	0.724
Ishida et al. 2012*	0.430	0.640	0.514
Richards et al. 2012*	0.724	0.654	0.687
Charnock et al. 2016	0.720	0.660	0.688
Lochner et al. 2016	0.900	0.840	**0.868**
This work (GBT)	0.829	0.860	**0.844**

Table 1. Test set performance comparison

Table 1 shows that our simple approach is comparable to the best method found in the literature (wavelet-based features and GBT).

By studying the position of the features on each tree, we can study how relevant a particular feature is for the classification. The most important feature is photo-z (1% of the relevance). Fig. 5 shows the importance of the feature vector. The most relevant features are located at the peaks of the *z* and *i* bands and in the decay-zone of the *r*-band. The *g*-band is the least important band.

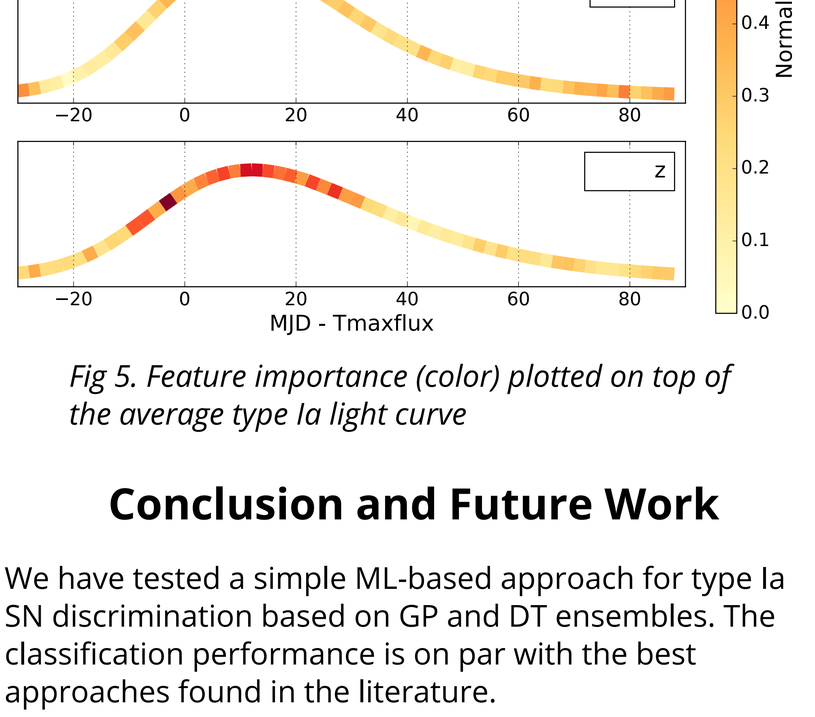

Fig 5. Feature importance (color) plotted on top of the average type Ia light curve

Conclusion and Future Work

We have tested a simple ML-based approach for type Ia SN discrimination based on GP and DT ensembles. The classification performance is on par with the best approaches found in the literature.

To improve this procedure we propose to

- Implement co-kriging GP to find correlations between bands
- Train the models with more sophisticated covariance matrices
- Try different sampling strategies for the feature vector based on the feature importance
- Incorporate other time series features
- Study in depth how the results depend on photo-z

As future work we plan on developing a domain-adaptation/transfer-learning strategy in order to use the models trained with SNPCC to classify other SNe datasets with different underlying distributions, such as CHASE and SUDARE

References

[1] P. Huijse et al, "Computational intelligence challenges and applications on large-scale astronomical time series databases." IEEE CIM 9.3, 2014
[2] R. Kessler et al, "Results from the supernova photometric classification challenge", PASP 122.898, 2010
[3] C. E. Rasmussen, "Gaussian processes for machine learning", MIT Press, 2006
[4] L. Mason et al. "Boosting algorithms as gradient descent in function space", NIPS, 1999
[5] L. Breiman, "Random Forests", Machine Learning 45, 2001
[6] J. W. Richards, et al, "Semi-supervised learning for photometric supernova classification", MNRAS 419.2, 2012
[7] E. Ishida and R.S. de Souza. "Kernel PCA for Type Ia supernovae photometric classification." MNRAS 430.1, 2013
[8] N. Karpenka, F. Feroz, and M. P. Hobson, "A simple and robust method for automated photometric classification of supernovae using neural networks", MNRAS, 2012
[9] M. Lochner et al, "Photometric Supernova Classification with Machine Learning", arXiv preprint, 2016
[10] Charnock and A. Moss, "Deep Recurrent Neural Networks for Supernovae Classification", arXiv preprint, 2016

Acknowledgements

This work was funded by CONICYT-CHILE under grant FONDECYT postdoctoral Nº 3150460, FONDECYT Nº 1140816 and CONICYT DPI 20140090. The authors acknowledge support from the Chilean Ministry of Economy, Development, and Tourism's Millennium Science Initiative through grant IC12009, awarded to The Millennium Institute of Astrophysics, MAS. Powered@NLHPC: This research was partially supported by the supercomputing infrastructure of the NLHPC (ECM-02)