Data Augmentation in a Hierarchical-Based Classification Scheme for Variable Stars

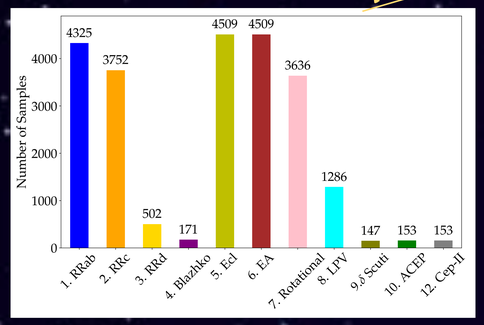
[1] [2] [3] [4]

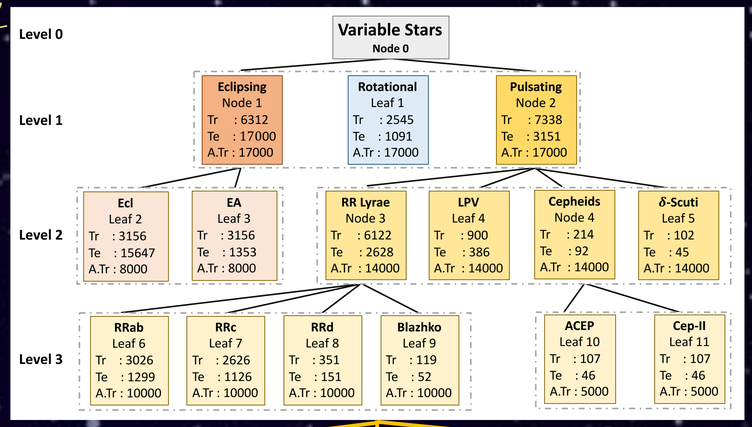
Zafiirah Hosenie[1], Robert Lyon[3], Benjamin Stappers[1], Arrykrishna Mootoovaloo[2] & Vanessa McBride[4]

1.Introduction

- A major issue that impedes the successful automated classification of astronomical data is the imbalanced learning problem.
- This problem impacts classification of variable stars in particular, as some types of variable stars are rare, making it difficult to build automated machine learning (ML) systems.

2.Hierarchical Taxonomy





Three methods of Data Augmentation:

1. **SMOTE: inserts artificially generated minority class examples into a data set by operating in 'feature space' rather than 'data space'**
2. **RASLE: *R*andomly *A*ugmented *S*ampled *L*ight curves from magnitude *E*rrors is employed on LCs or time series data directly. For each data point at a specific time, we sample a single magnitude from a probability distribution function (*pdf*) as going over the magnitude space vertically.**
3. **GPFit: We build a model describing variable stars using Gaussian Processes (GPs). We then randomly sampled synthetic LCs from the GP model to form the augmented training set.**

3.Methodology & Results

Stage 1: hierarchical tree classifiers

- We use the astrophysical properties of the various sources to construct a hierarchical-based structure to represent the different classes.
- For the HC, we use XGBoost and Random Forest.

Augmentation methods	Classifiers	G-Mean	Balanced-accuracy
First Level: Eclipsing, Rotational and Pulsating Classification			
H19 (No augmentation)	RF	0.78/0.78/0.86 (∼ 0.79)	0.59/0.60/0.75 (∼ 0.61)
SMOTE	XGBoost	0.80/0.77/0.89 (∼ 0.81)	0.63/0.59/0.80 (∼ 0.65)
	RF	0.80/0.78/0.89 (∼ 0.81)	0.63/0.60/0.79 (∼ 0.65)
RASLE	XGBoost	0.82/0.76/0.89 (∼ 0.83)	0.66/0.57/0.79 (∼ 0.68)
	RF	0.82/0.77/0.89 (∼ 0.83)	0.66/0.58/0.79 (∼ 0.68)
GpFit	XGBoost	0.80/0.75/0.89 (∼ 0.81)	0.63/0.56/0.79 (∼ 0.65)
	RF	**0.80/0.75/0.89 (∼ 0.81)**	**0.63/0.56/0.78 (∼ 0.65)**
Second Level: RR Lyrae, LPV, Cepheids and δ-Scuti			
H19 (No augmentation)	RF	0.99/1.00/0.97/1.00 (∼ 0.99)	0.98/0.99/0.93/1.00 (∼ 0.98)
SMOTE	XGBoost	0.99/1.00/1.00/0.95 (∼ 0.99)	0.97/0.99/1.00/0.90 (∼ 0.97)
	RF	0.99/1.00/1.00/0.96 (∼ 0.99)	0.97/0.99/1.00/0.92 (∼ 0.97)
RASLE	XGBoost	0.99/1.00/0.99/0.93 (∼ 0.99)	0.98/1.00/0.98/0.85 (∼ 0.98)
	RF	0.99/1.00/1.00/0.94 (∼ 0.99)	0.98/0.98/1.00/0.88 (∼ 0.98)
GpFit	XGBoost	0.99/1.00/0.99/0.95 (∼ 0.99)	0.97/0.99/0.97/0.99 (∼ 0.98)
	RF	**0.99/1.00/1.00/0.97 (∼ 0.99)**	**0.97/0.99/1.00/0.93 (∼ 0.98)**
Second Level: Ecl and EA			
H19 (No augmentation)	RF	0.93/0.93 (∼ 0.93)	0.86/0.86 (∼ 0.86)
SMOTE	XGBoost	0.94/0.94 (∼ 0.94)	0.88/0.88 (∼ 0.88)
	RF	0.94/0.94 (∼ 0.94)	0.88/0.88 (∼ 0.88)
RASLE	XGBoost	0.93/0.93 (∼ 0.93)	0.85/0.85 (∼ 0.85)
	RF	0.93/0.93 (∼ 0.93)	0.85/0.86 (∼ 0.86)
GpFit	XGBoost	0.93/0.93 (∼ 0.93)	0.88/0.88 (∼ 0.88)
	RF	**0.94/0.94 (∼ 0.94)**	**0.87/0.88 (∼ 0.88)**
Third Level: RRab, RRc, RRd and Blazhko			
H19 (No augmentation)	RF	0.97/0.92/0.65/0.44 (∼ 0.92)	0.94/0.85/0.40/0.18 (∼ 0.86)
SMOTE	XGBoost	0.95/0.92/0.67/0.58 (∼ 0.91)	0.91/0.83/0.42/0.31 (∼ 0.83)
	RF	0.95/0.82/0.47/0.33 (∼ 0.91)	0.91/0.82/0.47/0.33 (∼ 0.83)
RASLE	XGBoost	0.96/0.95/0.56/0.53 (∼ 0.92)	0.93/0.89/0.30/0.26 (∼ 0.87)
	RF	0.97/0.95/0.52/0.52 (∼ 0.92)	0.94/0.90/0.25/0.25 (∼ 0.87)
GpFit	XGBoost	0.97/0.93/0.57/0.44 (∼ 0.92)	0.94/0.86/0.30/0.17 (∼ 0.85)
	RF	**0.97/0.93/0.56/0.41 (∼ 0.92)**	**0.94/0.87/0.32/0.26 (∼ 0.87)**
Third Level: ACEP and Cep-II			
H19 (No augmentation)	RF	0.90/0.90 (∼ 0.90)	0.82/0.80 (∼ 0.81)
SMOTE	XGBoost	0.88/0.88 (∼ 0.88)	0.78/0.76 (∼ 0.77)
	RF	0.88/0.88 (∼ 0.88)	0.78/0.76 (∼ 0.77)
RASLE	XGBoost	0.88/0.88 (∼ 0.88)	0.77/0.78 (∼ 0.77)
	RF	0.88/0.88 (∼ 0.88)	0.77/0.78 (∼ 0.78)
GpFit	XGBoost	0.88/0.88 (∼ 0.88)	0.78/0.78 (∼ 0.78)
	RF	**0.91/0.91 (∼ 0.91)**	**0.84/0.82 (∼ 0.83)**

Stage 2: level-wise data augmentation in HC

- Since the training set is still imbalanced after aggregating subclasses into superclasses, we use the three data augmentation techniques: SMOTE, RASLE, GPFit

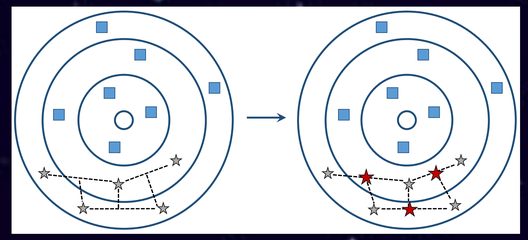
Stage 3: feature extraction

- Our features are based on 6 intrinsic statistical properties relating to location (mean magnitude), scale (standard deviation), variability (mean variance), morphology (skew, kurtosis, amplitude), and time (period).

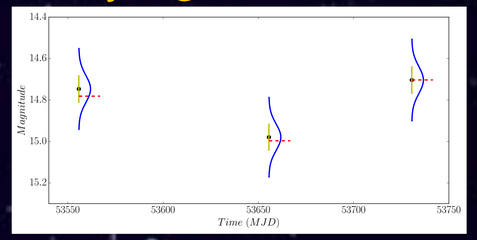
Stage 4: training with Bayesian optimization

- The training set moves through the first level in the HC scheme. The training examples are then augmented. The model (see dotted square at level 1) is trained using either RF or XGBoost classifier.
- Lastly, we evaluate our trained model on real LCs in the test set. The same concepts apply at different levels in the HC where real LCs move down the node, get augmented and classified in their respective classes.

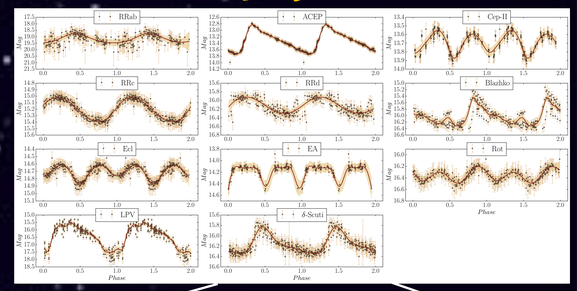
SMOTE

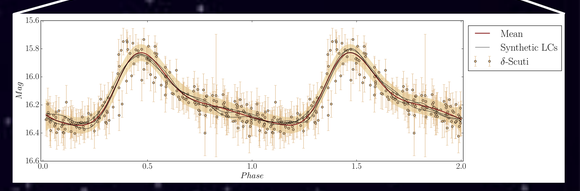


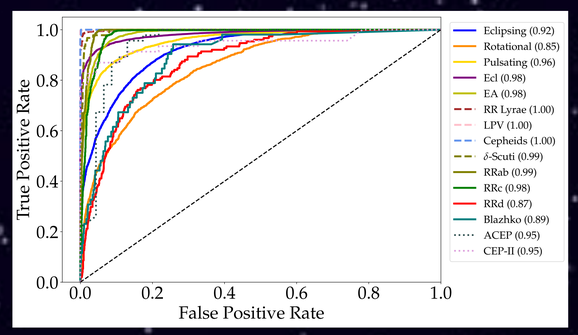
RASLE



GPFit







4. Conclusion

- When using GpFit method, our RF implementation performs best at all HC levels when compared to H19.
- We found that both XGBoost and RF provide good performance for variable star classification.
- We assess the consistency of the results using GpFit and RF by plotting the Receiver Operator Characteristic (ROC) curve for each class.