Two Statistical Challenges in Classification of Variable Sources

James Long

Texas A&M University

March 21, 2017

ICTS "Time Series Analysis for Synoptic Surveys and Gravitational Wave Astronomy"

Outline

Background on Automated Variable Star Classification

Example: CART Classifier Applied to OGLE

Challenge 1: Controlling Computational Costs in Feature Extraction

Challenge 2: Post Classification Inference

Outline

Background on Automated Variable Star Classification

Example: CART Classifier Applied to OGLE

Challenge 1: Controlling Computational Costs in Feature Extraction

Challenge 2: Post Classification Inference

Overview of Statistical Classification

Key Terms:

- **training data:** lightcurves of known class
- **unlabeled data:** lightcurves of unknown class

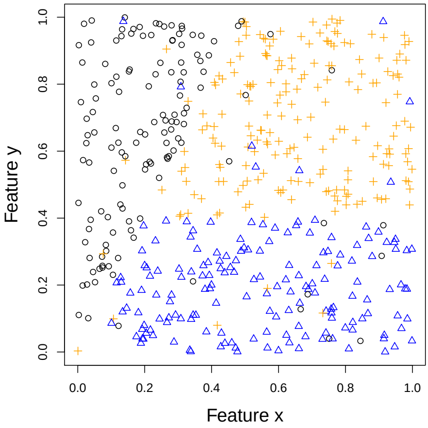
Steps in Classification:

1. **feature extraction:** derive quantities from light curves useful for separating classes, eg period, amplitude, derivatives, etc.
2. **classifier construction:** using training data, construct function

$$\widehat{\mathcal{C}}(features) \rightarrow class$$

3. **apply classifier:** for unlabeled data, compute features and predict class using $\widehat{\mathcal{C}}$

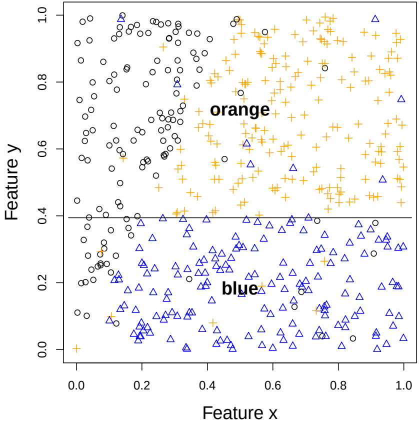
Classifier Construction using CART

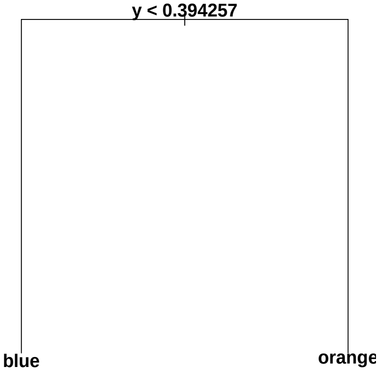


- Classification and Regression Trees (CART) developed in 1980s
- recursively partitions feature space
- partition represented by tree

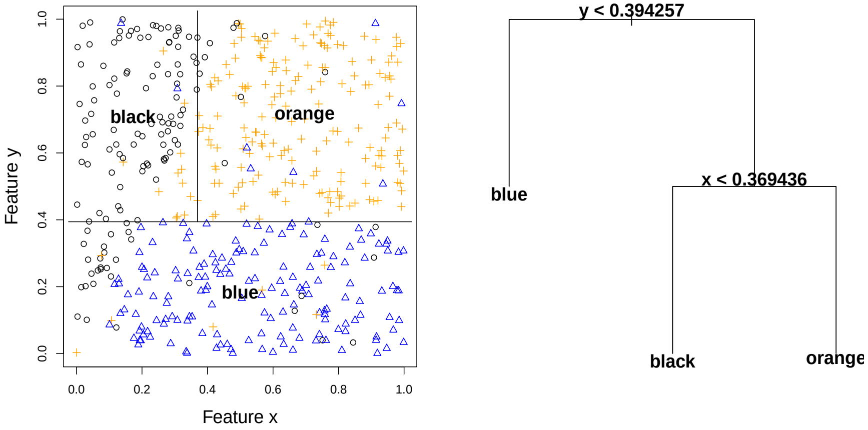
Breiman et al. 1984 "Classification and regression trees"

Building CART Tree . . .

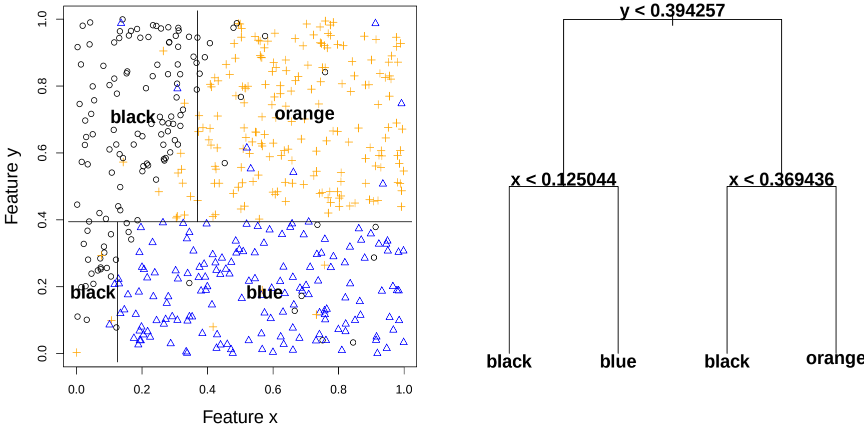




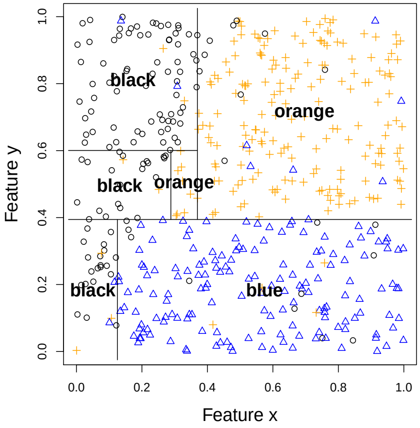
Building CART Tree . . .

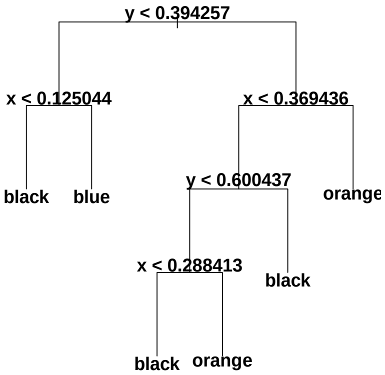


Building CART Tree . . .



Resulting Classifier





Apply Classifier to Test Data

Test Data: Data used to evaluate classifier accuracy. Test data is not used to construct classifier.

Confusion Matrix: Rows are true class of test data. Columns are predicted class of test data. Entries are counts.

	Predicted		
Truth	black	blue	orange
black	23	1	7
blue	2	30	2
orange	3	1	31

Outline

Background on Automated Variable Star Classification

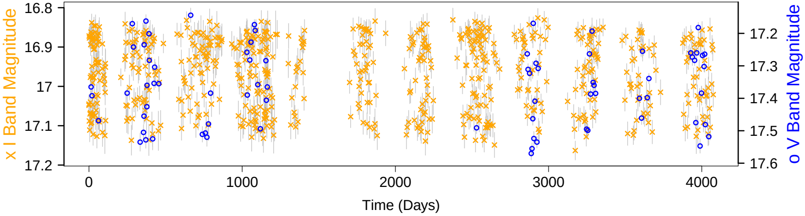
Example: CART Classifier Applied to OGLE

Challenge 1: Controlling Computational Costs in Feature Extraction

Challenge 2: Post Classification Inference

Optical Gravitational Lensing Experiment (OGLE)

- 400,000 + variable sources in LMC, SMC, Galactic Bulge
- typically hundreds of epochs in I, dozens in V
- 10 year + baseline



OGLE Variable Source Catalog: `http://ogledb.astrouw.edu.pl/~ogle/CVS/`

OGLE Classification Example

Classes

- Mira O–rich
- Mira C–rich
- Cepheid
- RR Lyrae AB
- RR Lyrae C

Features

- period (of best fitting sinusoid)
- amplitude $= 95^{th}$ percentile mag - 5^{th} percentile mag
- skew of magnitude measurements
- p2p_scatter[1]

[1] Dubath et al. 2011 "Random forest automated supervised classification of Hipparcos periodic variable stars" MNRAS

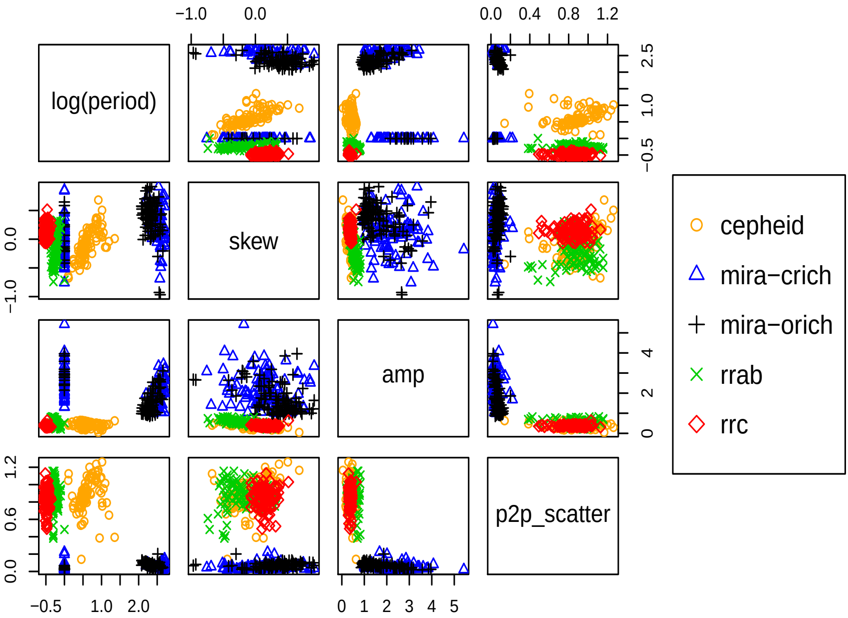
First 6 Rows of Feature–Class Dataframe

period	skew	amp	p2p_scatter	class
1.6128497	-0.5009063	0.56050	0.8672024	cepheid
0.6394983	0.3022388	0.35675	0.7523166	rrab
0.6433533	0.3200730	0.33730	0.8554517	rrab
0.4954661	-0.2053132	0.42000	0.7560226	rrab
0.3540801	0.1361693	0.34340	0.9215426	rrc
0.5460332	-0.3863142	0.69600	1.0682803	rrab

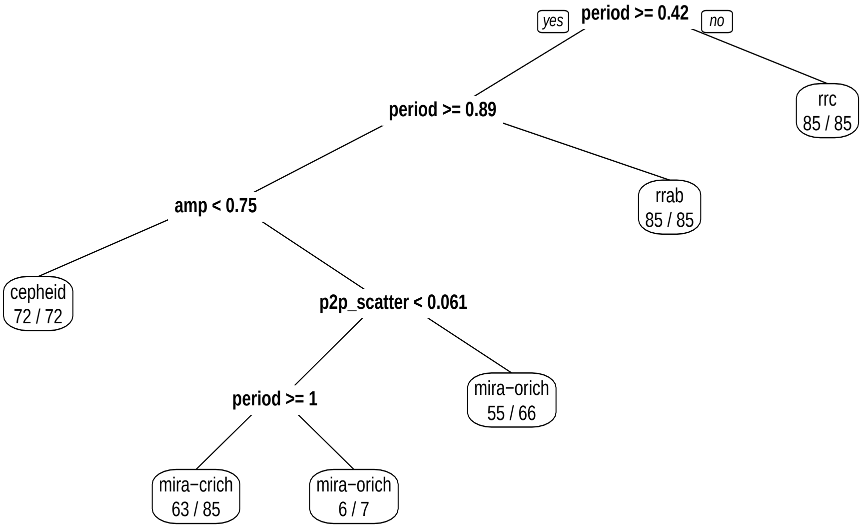
500 total rows. 5 classes.

training data: 400 randomly selected rows
test data: remaining 100 rows

Feature Distributions



CART Model Fit To Training Data



Confusion Matrix using Test Data

	Predicted				
Truth	cepheid	mira-crich	mira-orich	rrab	rrc
cepheid	24	0	0	0	0
mira-crich	0	15	10	0	0
mira-orich	0	5	12	0	0
rrab	1	0	0	14	0
rrc	0	0	0	1	14

Conclusion: Develop features to better separate O/C–rich Mira.

Note: CART is interpretable (not black box) but not particularly accurate. Forms basis for Random Forests.[2]

[2] Breiman 2001. "Random forests" *Machine learning*

Outline

Background on Automated Variable Star Classification

Example: CART Classifier Applied to OGLE

Challenge 1: Controlling Computational Costs in Feature Extraction

Challenge 2: Post Classification Inference

Features and Computation Time

feature	computation time / l.c.
colors	≈ 0
Stetson–J[3]	≈ 0
period (best fitting sine)[4]	5 seconds
Mira Gaussian Process model[5]	20 sec
RR Lyrae template goodness–of–fit[6]	30 minutes
generative model posterior probabilities	ask David Jones
$\vdots$	$\vdots$

Computational limitations prevent extracting all features for all sources.

[3] Stetson 1996 "On the automatic determination of light-curve parameters for cepheid variables" PASP
[4] Vanderplas 2015 "Periodograms for multiband astronomical time series" ApJ
[5] He 2016 "Period Estimation for Sparsely Sampled Quasi–periodic Light Curves Applied to Miras" ApJ
[6] Sesar 2016 "Machine–learned Identification of RR Lyrae Stars from Sparse, Multi–band data: The PS1 Sample"

Minimizing Feature Computations

Common Solution:

1. compute cheap features for all sources
2. build a simple classifier
3. select "interesting objects"
4. compute more expensive features on interesting objects, build classifier

Example: Variable versus non–variable

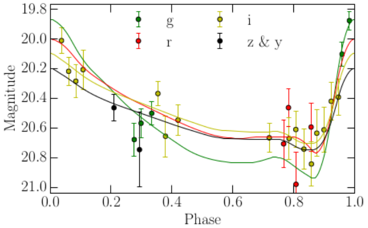
1. compute Stetson J, other variability metrics
2. make cuts on variability metrics
3. compute more expensive features on objects classified as variables

Multiple Iterations: RR Lyrae in Pan–STARRS

Example: Sesar 2016

Goal: Find RR Lyrae among 500 million Pan–STARRS objects

- Classifier 1: identified variables using Stetson J, other metrics
- Classifier 2: extracted "simple" features (multiband period estimator, amplitude, etc.) on variables, built classifier
- Classifier 3: extracted computationally intensive features (eg RRL template fits) on high probability RRL candidates from Classifier 2, built classifier



Sesar 2016 "Machine–learned Identification of RR Lyrae Stars from Sparse, Multi–band data: The PS1 Sample"

Formalizing this Framework

Standard Setting

- l is light curve
- $f(l) = X$ is features for light curve l
- Z is class of l
- $\widehat{C}$ is classifier

Train classifier $\widehat{C}$ to

$$\text{maximize } P(\widehat{C}(f(l)) = Z)$$

Controlling Feature Extraction Computational Cost

- classifier $\widehat{C}$ <u>chooses</u> which features to compute
- $\widehat{C}$ outputs predicted class $\widehat{Z}$ and feature extraction time T

$$C(l) = (\widehat{Z}, T)$$

Train classifier $\widehat{C}$ to

$$\text{maximize } P(\widehat{C}(l)_1 = Z)$$
$$\text{subject to } \mathbb{E}[\widehat{C}(l)_2] < t_0$$

Result

$$\approx Nt_0 \text{ time to classify } N \text{ objects}$$

Question: Has this been studied in the statistics / ML literature?

Outline

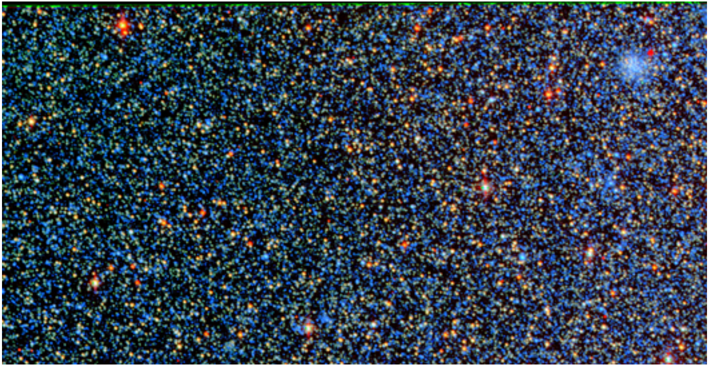
Background on Automated Variable Star Classification

Example: CART Classifier Applied to OGLE

Challenge 1: Controlling Computational Costs in Feature Extraction

Challenge 2: Post Classification Inference

What are the distances to these objects?



Problem:

$$\text{brightness} \propto \frac{\text{luminosity}}{\text{distance}^2}$$

Only brightness can be directly measured.

Image Source: DES Collaboration

Standard Candles

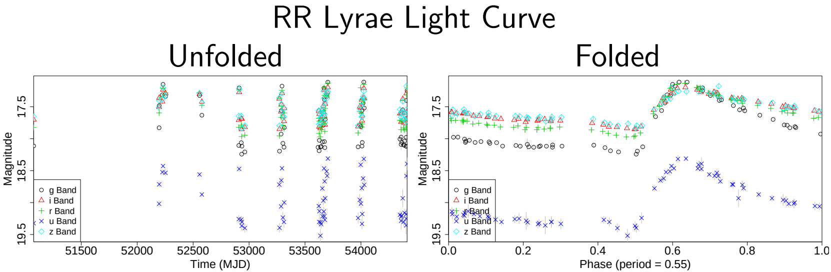
Standard Candle: Class of objects with same luminosity

- Know absolute luminosity of standard candle.
- Determine object is standard candle and estimate its brightness.
- Solve for distance.

RR Lyrae (RRL): Standard candle variable star

- All RR Lyrae have (approximately) same luminosity

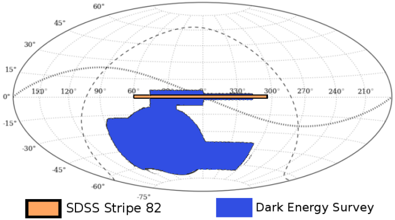
RR Lyrae are Variable Stars



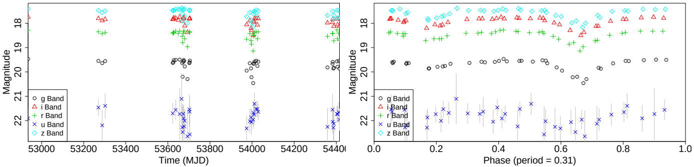
Standard candle: Distance to this star is proportional to mean magnitude, after accounting for dust and PL relation.

Sloan Digital Sky Survey (SDSS) III – Stripe 82

- Discovered $\approx 60,000$ variable stars
- ≈ 250 brightness measurements / star
- variables belong to many **classes**



Example Light Curve: Eclipsing Binary (Unfolded and Folded)



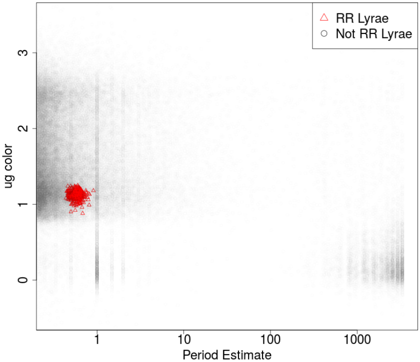
Ivezic 2007 "Sloan Digital Sky Survey Standard Star Catalog for Stripe 82: The Dawn of Industrial 1% Optical Photometry" ApJ.

Identifying RRL, Mapping MW Halo with SDSS

Sesar 2010:

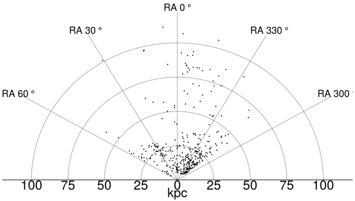
1. extracted features for $\approx 60,000$ variables (eg period, amplitude)
2. identified ≈ 350 RR Lyrae
3. estimated distances to RRL

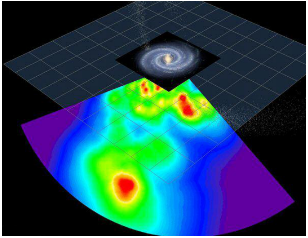
Steps 1 and 2



Sesar 2010 ApJ

Step 3

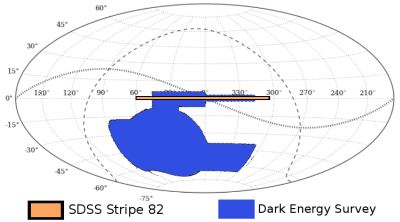


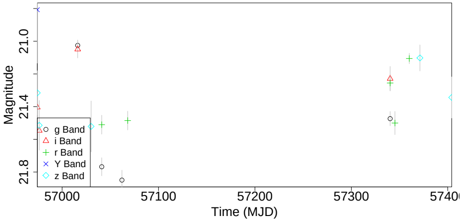


Mapping the Galactic Halo with DES

Dark Energy Survey (DES)

- ongoing survey (started 2013, 5 years of planned observing)
- 5000 square degrees ($\approx 1/9^{th}$ entire sky)
- depths to 24 mag in i
- 68 million stars
- $\approx$ 10 observations in each filter (g,i,r,z,Y) over five years





DES is deeper and wider but sparsely sampled.

See: Sesar 2016 "Machine–learned Identification of RR Lyrae Stars from Sparse, Multi–band data: The PS1 Sample" for similar work using Pan–STARRS

Complicated, Multilevel Inference Process

Steps in Inference Process:

1. classify stars as RR Lyrae
2. estimate distances to stars classified as RR Lyrae
3. estimate intensity maps of distribution of matter in MW halo

Can machine learning methods propagate uncertainty through all of these steps?

A Framework for Statistical Inference in Astrophysics

Chad M. Schafer

Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213;
email: cschafer@cmu.edu

Discusses multistage aspect of several astrostatistics problems.

Thank you. Questions?