New Techniques in Light Curve Analysis

Joseph Richards

UC Berkeley
Department of Astronomy
Department of Statistics
jwrichar@stat.berkeley.edu

High Energy Astrophysics Division Meeting

September 7, 2011

Center for Time-Domain Informatics

UC Berkeley (UCB):

Faculty/Staff

Josh Bloom, Dan Starr (Astro), John Rice, Noureddine El Karoui (Stats), Martin Wainwright, Masoud Nikravesh (CS)

Post-Docs

Dovi Poznanski, Brad Cenko, Nat Butler, Berian James, JWR

Grad Students

Dan Perley, Adam Miller, Adam Morgan, Chris Klein, James Long, Tamara Broderick, Sahand Negahban, John Brewer, Henrik Brink, Sharmo Bhattacharyya

Undergrads

Maxime Rischard, Justin Higgins, Rachel Kennedy, Jason Chu, Arien Crellin-Quick, Pierre Christian, Tatyana Gavrilchenko, Stuart Gegenheimer, Anthony Paredes, Benjamin Gerard

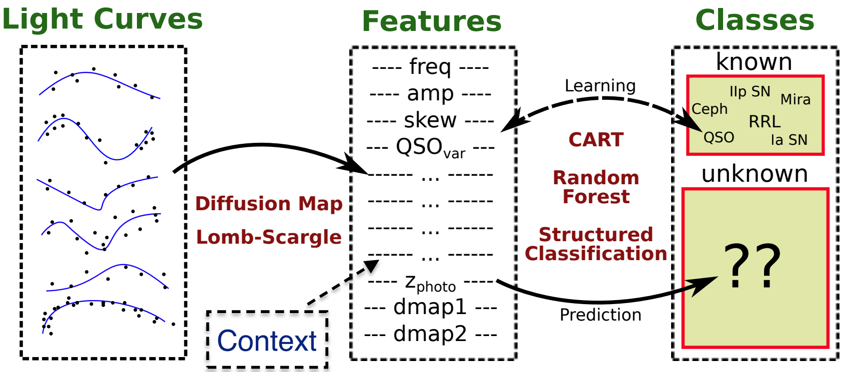
Lawrence Berkeley National Laboratory (LBNL):

Peter Nugent, David Schlegel, Nic Ross, Horst Simon

Visit our website: http://cftd.info/

Motivation

A road map for light curve classification



See: Richards et al. (2011) arXiv:1101.1959
Bloom & Richards (2011) arXiv:1104.3142

Motivation: Automated Learning on Light Curves

Need **machine learned classification** of light curves for:

1. detection and discovery of events in real time, condensing a data deluge into a trickle of astrophysical goodness
2. optimal allocation of (expensive!) follow-up resources, often in real time
3. construction of pure & complete samples of, e.g., Type Ia Supernovae (expansion history of Universe), RR Lyrae Variable Stars (structure of Milky Way), Eclipsing star systems (stellar mass, radius, age, distance)
4. outlier detection to find objects from new or rare classes

   Bhattacharyya et al. (2011) in prep.: semi-supervised anomaly detection

Discovery on massive data streams is not assured!

Example: Optimal Resource Allocation

Problem statement

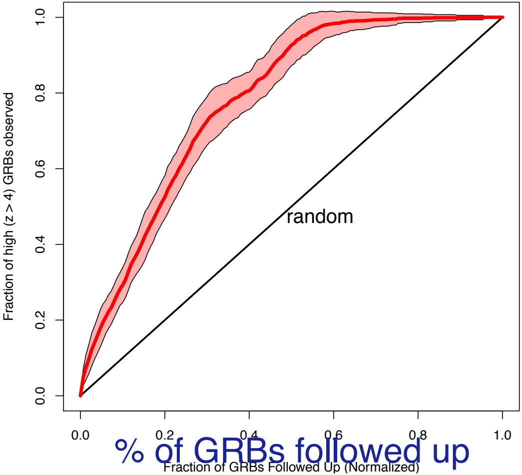
Given limited follow-up time, maximize the time spent on high-redshift GRBs

Based only on early-time metrics

Classification drives resource allocation

`RATE-GRBz`: **web tool for GRB follow-up**

Classification Efficiency



Morgan et al. 2011, in prep.

Machine-Learned Classification of Light Curves

with Josh Bloom, Dan Starr, Nat Butler, Darren Homrighausen, Chad Schafer, Peter Freeman, Dovi Poznanski

Bloom & Richards (2011) arXiv:1104.3142 - Overview of ML LC Class.
Richards et al. (2011) arXiv:1101.1959 - VarStar Classification
Richards et al. (2011) arXiv:1103.6034 - SN Typing
Bloom, et al. (2011) arXiv:1106.5491 - Classification for PTF

Light Curve Features

Domain knowledge drives choice of features

Periodic Metrics

Use generalized Lomb-Scargle method to find frequencies, amplitudes, phase offsets of fundamental freqs and harmonics

Variability Metrics

- Stetson indices
- damped random walk

QSO model of Butler & Bloom 2011

- point-to-point metrics

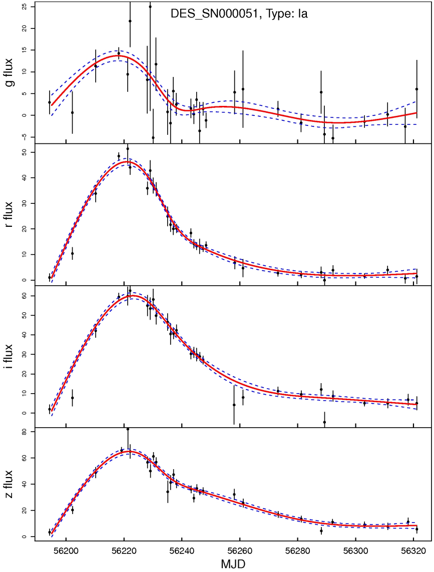
Shape Analysis

- marginals: std, skewness, kurtosis, ratios of quantiles
- **Low-D embeddings of LCs (e.g., diffusion map, LLE)**

Context Features

e.g., distance to nearest galaxy, type of nearest galaxy, location in the ecliptic plane, SDSS, etc.

Diffusion Map for Photometric SN Typing



Diffusion map – non-linear method to uncover low-dimensional structure in data (Lafon & Lee 2006)

- Map each light curve, $\mathbf{x}$, into m-dimensional diffusion space
  $\mathbf{x} \mapsto \{\psi_1(\mathbf{x}), \ldots, \psi_m(\mathbf{x})\}$
- Features for classification are the diffusion map coordinates

Richards et al. (2011) arXiv:1103.6034

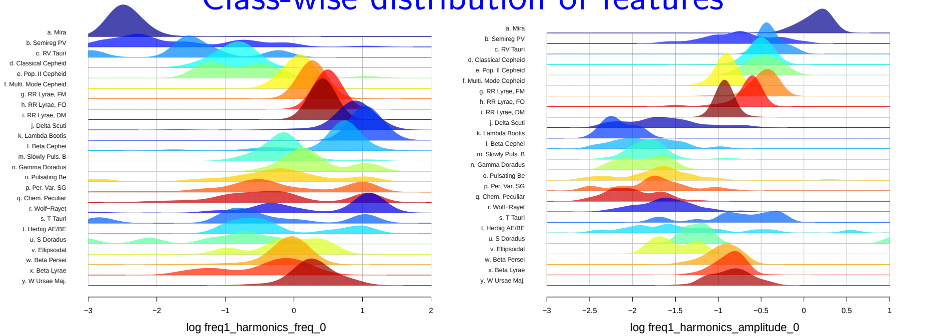
From Features to Classification

Classification:

We describe each light curve with a vector of features, $\mathbf{x}$

Goal: Using known labels $y_1, \ldots, y_n$, estimate model $\widehat{f}(\mathbf{x})$ to predict class probabilities for new light curves

Class-wise distribution of features



Frequency

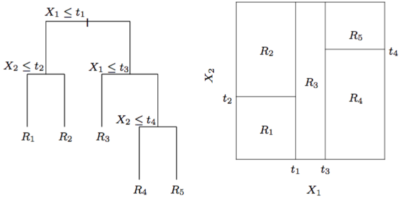
Amplitude

Classification: Decision Trees

Classification: Learn model $\widehat{f}(\mathbf{x})$ that maps a feature vector $\mathbf{x}$ to a vector of class probabilities.

Classification Trees:

- Binary partitions of feature space
- Each split minimizes node impurity
- Within each node, model class probabilities, $\widehat{f}(\mathbf{x})$, as constant



Hastie, Tibshirani, Friedman (2009)

Advantages:

1. Able to capture complex interactions
2. Robust to outliers
3. Handle multi-class problems
4. Immune to irrelevant features
5. Cope with missing values
6. Computationally efficient & scalable

Classification: Ensemble Methods

Drawback of Classification Trees

Classification trees are usually unbiased if grown deep enough, but have high variance

Note: Expected classification error is variance plus bias-squared

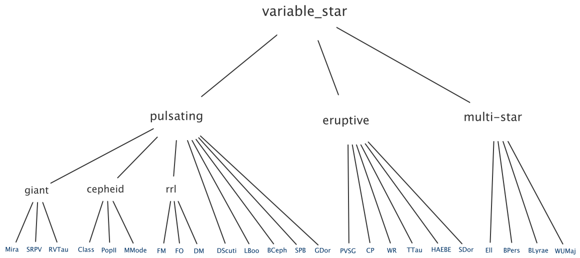
- **Bagging** averages trees from bootstrapped versions of $\mathbf{x}$
- **Boosting** averages a series of trees, iteratively up-weighting mis-classified data
- **Random Forest** averages B **de-correlated**, bootstrapped trees, $\widehat{f}_{\mathrm{RF}} = \frac{1}{B}\sum_{i=1}^{B} \widehat{f}_i$.

$$Var(\widehat{f}_{\mathrm{RF}}) = \rho Var(\widehat{f}_i) + \frac{1-\rho}{B} Var(\widehat{f}_i)$$

where ρ is the correlation between trees, $\widehat{f}_i$.

Classification: Structured Classification

Idea: Let class taxonomy guide classifier



HSC: Hierarchical single-label classification.

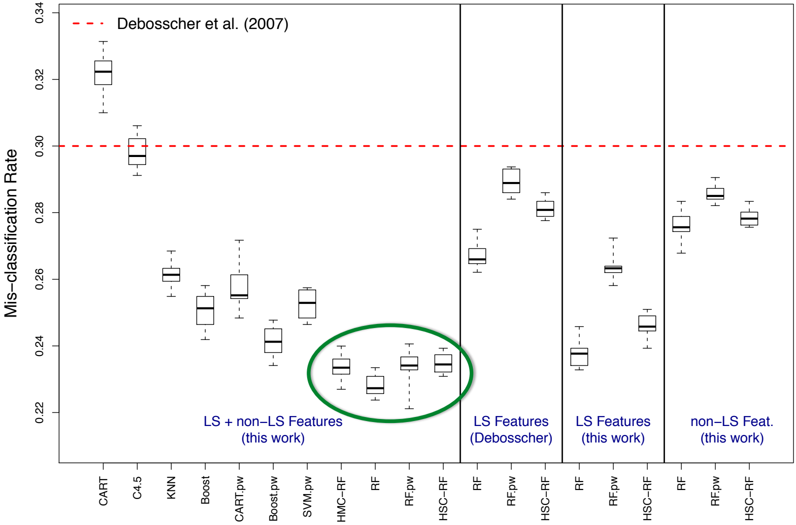
- Fit separate classifier at each non-terminal node.

HMC: Hierarchical multi-label classification.

- Fit one classifier, where $L(y, \widehat{f}(\mathbf{x})) \propto w_0^{\mathrm{depth}}$

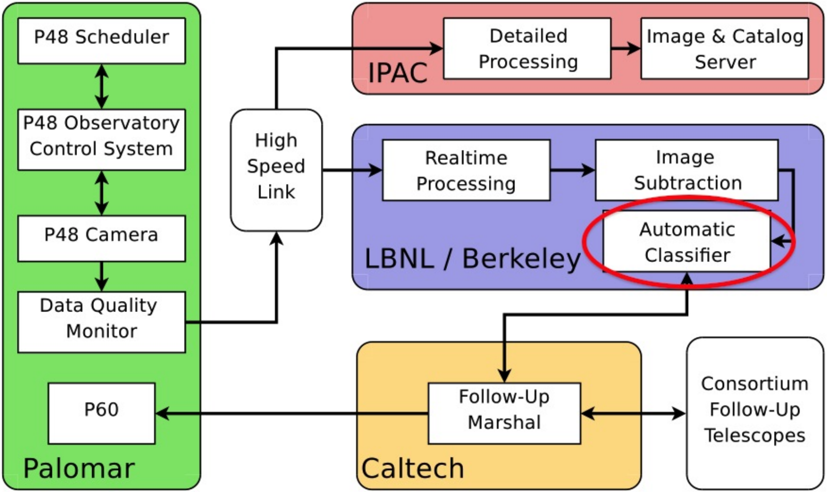
Classification of Hipparcos + OGLE VarStars

Cross-validated classification error rates



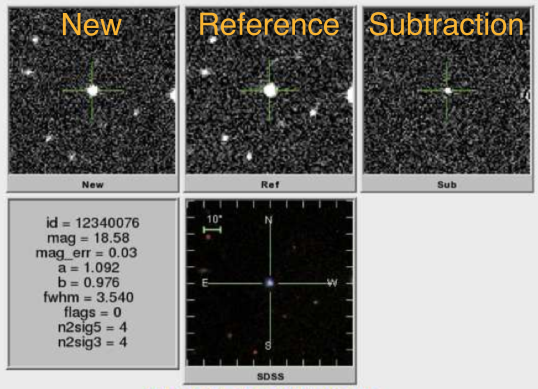
Classification for Palomar Transient Factory

Law et al. (2009, PASP, 121, 1395)



Classification for Palomar Transient Factory

Is this detection a real astrophysical source?



Negahban, et al. (2011), in prep.

PTF obtains 1.5M detections per night

Only 0.1% are real astrophysical sources!

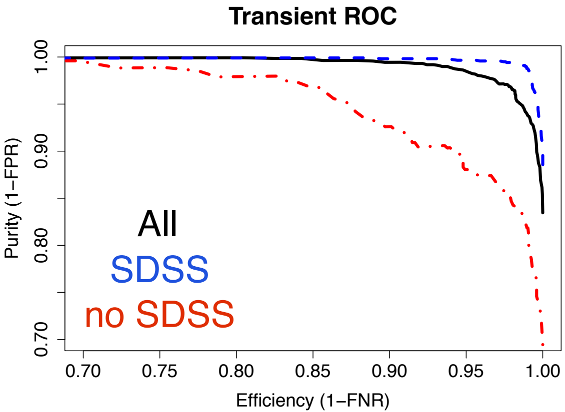
RF RB2 Classifier

Obtain $\sim$ 15% missed detection rate at 99% purity

Recently discovered SN2011fe, the most nearby SN found in the last $\sim$40 years

Classification for Palomar Transient Factory

Classification of newly discovered sources at time of discovery!



Bloom et al. (2011) arXiv:1106.5491

Random Forest classifier with **context** and **light curve features**

99.7% transient classification efficiency at 90% purity

Automated classifier drives follow-up!

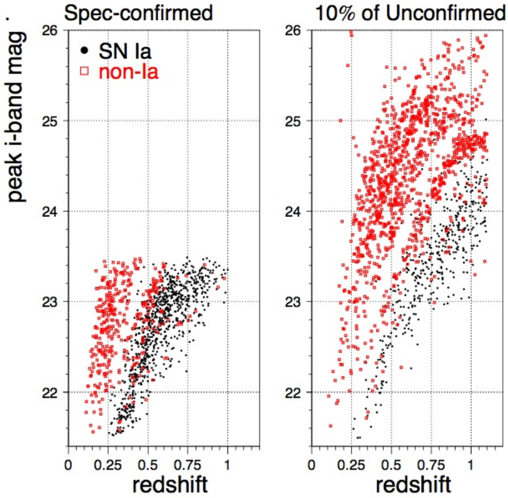
Sample Selection Bias in Light Curve Classification

with Dan Starr, Adam Miller, Nat Butler, James Long, John Rice, Josh Bloom (UC Berkeley), Henrik Brink & Berian James (DARK)

Richards et al. (2011), arXiv:1106.2832

Sample Selection Bias

In astronomical problems, the training (labeled) and testing (unlabeled) sets are often generated from different distributions.



Left: Training set
Right: Testing set

This problem is referred to as Sample Selection Bias or Covariate Shift.

SN Challenge Data

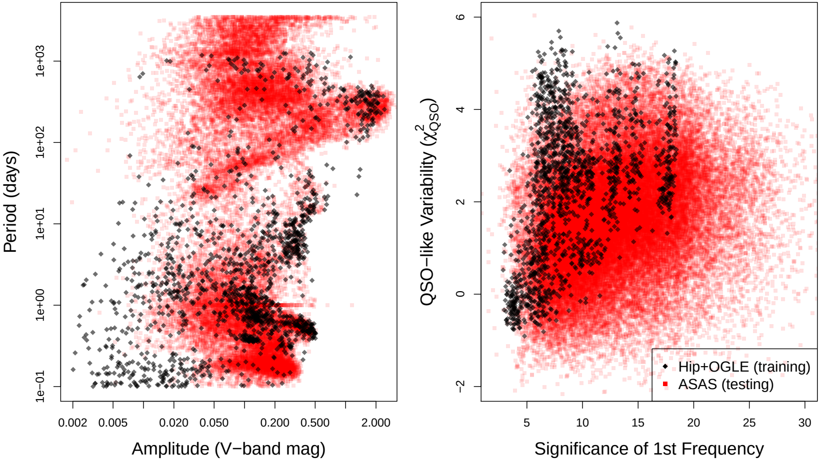
Kessler et al. (2010)

arXiv:1008.1024

Sample Selection Bias: VarStar Classification

Black: Training set (OGLE+Hipparcos, see Debosscher et al. 2007)

Red: Testing set (All Sky Automated Survey, ASAS; Pojmanski 2002)



Sample Selection Bias in Astronomy Datasets

Training sets in astronomy are biased:

1. Populations of well-studied objects are inherently biased toward brighter/nearby sources with better quality data
2. Available training data are typically from older, lower quality detectors
3. Each survey has different characteristics, aims, cadences...
4. Training data are often generated from idealized models

This can cause significant problems for off-the-shelf supervised methods:

1. Poor model selection – risk minimization (e.g., by cross-validation) is performed with respect to $\mathbf{P}_{\mathrm{Train}}(\mathbf{x}, y)$
2. Regions of feature space ignored by the training data – catastrophically bad extrapolation

Methods: Active Learning (AL)

Active Learning: Identify and manually label the testing set data that would most help future iterations of the classifier

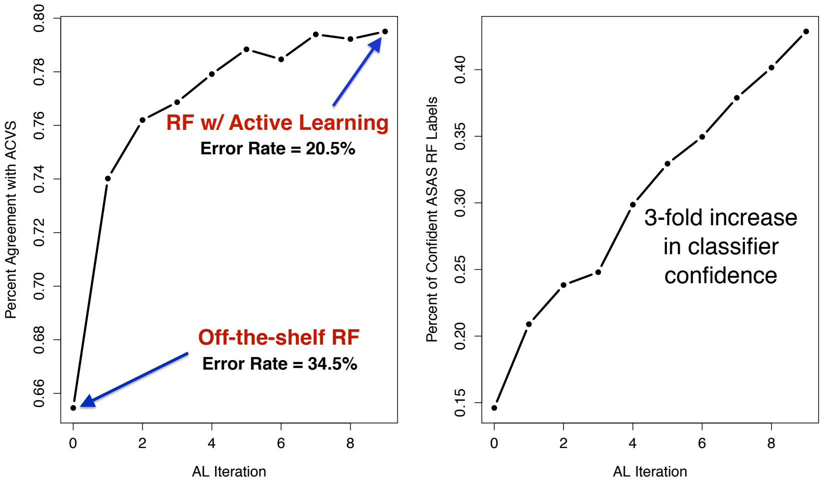
Key: In astronomy, we often have the ability to selectively follow up on sources:

- Spectroscopic study
- Query other databases; cross-match
- "Look at" the data; Citizen Science projects

Pool-based, batch-mode Active Learning: On each AL iteration, select a batch of objects from the entire testing set for manual labeling via a **query function**

Results: All Sky Automated Survey (ASAS)

Performance metrics of classifier vs. AL iteration:



from Richards et al. (2011), arXiv:1106.2832

Summary

- Machine learning is crucial for time-domain surveys
  - Methods & algorithms that handle large data rates
  - Statistical guarantees on performance
  - Reproducible and transparent!
  - **Both astrophysical insight and machine learning expertise are essential elements in this endeavor!**
- Some of our ongoing research for LC analysis
  1. Period estimation methods.
  2. Techniques to automatically extract low-D structure: LLE, diffusion map, etc.
  3. Structured classification to exploit taxonomy
  4. Active learning to overcome sample selection bias
  5. Noisification & de-noisification approaches to analyze low S/N data (Long et al., in prep)
  6. Semi-supervised anomaly detection

Center for Time-Domain Informatics Publications

Starr, D. L., Bloom, J. S., Brewer, J. M., Butler, N. R., Poznanski, D., Rischard, M., Klein, C. **The Berkeley Transient Classification Pipeline: Deriving Real-time Knowledge from Time-domain Surveys** (2009, ASPC, 411, 493)

Butler, Nathaniel R., Bloom, Joshua S. **Optimal Time-Series Selection of Quasars** (2011, AJ, 147, 93)

Richards, Joseph W., et al. **On Machine-Learned Classification of Variable Stars with Sparse and Noisy Time-Series Data** (2011, ApJ, 733, 1)

Bloom, Joshua S. & Richards, Joseph W. **Data Mining and Machine-Learning in Time-Domain Discovery & Classification** (2011, Chapter in the forthcoming book "Advances in Machine Learning and Data Mining for Astronomy")

Richards, Joseph W., Homrighausen, Darren, Freeman, Peter E., Schafer, Chad M. & Poznanski, Dovi **Semi-supervised Learning for Photometric Supernova Classification** (2011, accepted, MNRAS)

Richards, Joseph W., et al. **Active Learning to Overcome Sample Selection Bias: Application to Photometric Variable Star Classification** (2011, arXiv:1106.2832)

Bloom, Joshua S., Richards, Joseph W., et al. **Automating Discovery and Classification of Transients and Variable Stars in the Synoptic Survey Era** (2011, arXiv:1106.5491)