Time Series Classification in the Synoptic Surveys Era

Joseph Richards
& Joshua Bloom

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

Statistical Challenges in Modern Astronomy V
June 16, 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: Light Curves

**Supernovae**
- Type Ia SN, DES simulation

**Variable Stars**
- Pulsating star, Hipparcos Survey
- Eclipsing binary system, OGLE

http://dotastro.org
Motivation: Zoo of Science Classes

Hundreds of classes of time-varying objects & events

Source: http://dotAstro.org
Motivation: Automated LC Classification

Need accurate **automated classification** of light curves for:

1. **optimal allocation of** (expensive!) **follow-up resources**, often in real time

2. **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)

3. **outlier detection** to find objects from new or rare classes.
A road map for light curve classification:

Light Curves

Features

- freq
- amp
- skew
- QSO_{var}
- ...
- ...
- ...
- ...
- ...
- z_{photo}
- dmap1
- dmap2

Classes

known
- IIp SN
- Mira
- Cepheid
- RRL
- QSO
- Ia SN

unknown

Diffusion Map

Lomb-Scargle

Context

Learning

CART
Random Forest
Structured Classification

Prediction

Ex: High-$z$ Gamma Ray Burst Classification

- **GRBs**: short-lived blasts of high energy light
- Two origins: exploding massive stars & colliding compact objects
- < 20 high-$z$ bursts have ever been discovered!
Motivation: 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

See poster by Adam Morgan
Motivation: Data Deluge from New Surveys

Large Synoptic Survey Telescope (LSST) - 2020
- Light curves for 800M sources every 3 days
- $10^6$ SNe/yr, $10^7$ eclipsing binaries
- 3.2 gigapixel camera, 20 TB/night!!

Gaia space astrometry mission - 2013
- 1 billion stars observed $\sim 70$ times over 5 years
- Will observe 20K supernovae

Many other astronomical surveys are already producing data: SDSS, PTF, Pan-STARRS, Hipparcos, OGLE, ASAS, Kepler, LINEAR, DES (soon) etc., etc.
Machine-Learned Classification of Light Curves

with Dan Starr, Nat Butler, Josh Bloom

**Variable Star Classification**

**Hipparcos**: Space mission. Precise astrometry.

**OGLE**: Las Campanas, Chile. Gravitational lensing.

<table>
<thead>
<tr>
<th>Survey</th>
<th>Hipparcos</th>
<th>OGLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_{LC}$</td>
<td>1044</td>
<td>523</td>
</tr>
<tr>
<td>$\bar{T}$ (days)</td>
<td>1097</td>
<td>1067</td>
</tr>
<tr>
<td>$\bar{N}_{epochs}$</td>
<td>103</td>
<td>329</td>
</tr>
</tbody>
</table>

Sources from 25 science classes

Data set compiled by Debosscher et al. (2007)
Features for VarStar Classification

We typically use 40-60 features for light curve classification:

**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)

**Context Features**
e.g., distance to nearest galaxy, type of nearest galaxy, location in the ecliptic plane, SDSS, etc.
Classification:

We describe each light curve with a set of 50+ features, $x$.

**Goal:** Using known labels $y_1, \ldots, y_n$, estimate model $\hat{f}(x)$ to predict class probabilities for new light curves.
Performance on Hipparcos + OGLE Data

Cross-validated classification error rates

Debosscher et al. (2007)

LS + non–LS Features
(this work)

LS Features
(Debosscher)

LS Features
(this work)

non–LS Feat.
(this work)
Classification for Palomar Transient Factory

Law et al. (2009, PASP, 121, 1395)
PTF: Real-Bogus Classification

Is this detection a real astrophysical source?

PTF obtains $\gtrsim 10^6$ detections per night

Only 0.1% are real astrophysical sources!

RF RB2 Classifier

Using subtraction image & context features, obtain $< 20\%$ missed detection rate at 99% purity

Is the discovered source a transient?

Classification at time of discovery!

Random Forest classifier using context features and crude light curve features at time of discovery.

99.7% transient classification efficiency at 90% purity.

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
In astronomical problems, the training (labeled) and testing (unlabeled) sets are often generated from different distributions. This problem is referred to as Sample Selection Bias or Covariate Shift.

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: SN Typing

For SN Ia typing, it is better to use deeper spectroscopic training samples, even though they produce data from fewer SNe.

$S_{m,25}$ - 25th mag-limited spec survey is optimal (23.5th mag was used in SN Challenge)

From Richards et al. (2011) arXiv:1103.6034

Figure: Type Ia SN Purity and Efficiency of diffusion map Random Forest classifier on SN Challenge testing data
**Black**: Training set (OGLE+Hipparcos, see Debosscher et al. 2007)

**Red**: Testing set (All Sky Automated Survey, ASAS; Pojmanski 2002)
Sample Selection Bias in VarStar Classification

Training sets in variable star studies are biased:

1. Populations of well-studied objects are inherently biased toward brighter/closer 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 $P_{\text{Train}}(x, y)$
2. Regions of feature space ignored by the training data – catastrophically bad extrapolation
### Example: ASAS varstar classification (50,124 stars in ACVS)

Results of off-the-shelf Random Forest classifier:

<table>
<thead>
<tr>
<th>Class</th>
<th>ACVS Class</th>
<th>RF Predicted Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Mira</td>
<td>0.91</td>
<td>0.01</td>
</tr>
<tr>
<td>d. Classical Cepheid</td>
<td>0.27</td>
<td>0.01</td>
</tr>
<tr>
<td>e. Pop. II Cepheid</td>
<td>0.04</td>
<td>0.01</td>
</tr>
<tr>
<td>f. Multi. Mode Cepheid</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>g. RR Lyrae, FM</td>
<td>0.81</td>
<td>0.05</td>
</tr>
<tr>
<td>h. RR Lyrae, FO</td>
<td>0.01</td>
<td>0.27</td>
</tr>
<tr>
<td>j. Delta Scuti</td>
<td>0.01</td>
<td>0.38</td>
</tr>
<tr>
<td>l. Beta Cephei</td>
<td>0.14</td>
<td>0.29</td>
</tr>
<tr>
<td>q. Chem. Peculiar</td>
<td>0.02</td>
<td>0.11</td>
</tr>
<tr>
<td>w. Beta Persei</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>x. Beta Lyrae</td>
<td>0.06</td>
<td>0.02</td>
</tr>
<tr>
<td>y. W Ursae Maj.</td>
<td>0.08</td>
<td>0.01</td>
</tr>
<tr>
<td>UNKNOWN</td>
<td>0.03</td>
<td>0.14</td>
</tr>
</tbody>
</table>

The table above shows the comparison between the ACVS classifications and the RF predictions for each class of varstars. The numbers represent the probability of each classification. The table highlights the cases where the ACVS and RF classifications do not match.
Some Methods
Methods: Importance Weighting (IW)

**Idea:** Choose classifier that minimizes statistical risk over distribution of the *testing* set.

- **Idea:** Choose classifier that minimizes statistical risk over distribution of the *testing* set.
Methods: Importance Weighting (IW)

**Idea:** Choose classifier that minimizes statistical risk over distribution of the *testing* set.

Use importance weights on *training* set:

$$w_i = \frac{P_{Test}(x_i, y_i)}{P_{Train}(x_i, y_i)} = \frac{P_{Test}(x_i)P_{Test}(y_i|x_i)}{P_{Train}(x_i)P_{Train}(y_i|x_i)} = \frac{P_{Test}(x_i)}{P_{Train}(x_i)}$$

**Issues:**

1. Difficult to estimate high-dimensional feature densities.
2. IW is asymptotically sub-optimal when the statistical model is correctly specified (Shimodaira 2000)
3. Requires the support of the testing distribution be a subset of the support of the training distribution
Methods: Co-training (CT)

Idea: Iteratively add to the training set the most confidently classified testing data.
Methods: Co-training (CT)

Idea: Iteratively add to the training set the most confidently classified testing data

CT approach (Blum & Mitchell 1998)

Iterate until all data are in training set:

1. Build two separate classifiers, $h_1$ & $h_2$ on disjoint feature sets $x_1$ & $x_2$

2. Add the most confidently classified testing instances to the training set of the other classifier

Final classifier: $p(y|x) = h_1(y|x_1)h_2(y|x_2)$

Self-training (ST) performs iterations on a single classifier

Drawback: CT & ST are greedy: dominant classes in the training data gain undue influence
**Methods: Active Learning (AL)**

**Idea:** Manually label the testing 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

On each AL iteration, select a batch of objects from the entire testing set for manual labeling via a query function (pool-based, batch-mode AL).

**Heuristic:** Query data in regions of feature space that are densely populated with testing data and sparsely populated with training data.
Methods: Active Learning (AL)

**Idea:** Manually label the testing 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

On each AL iteration, select a batch of objects from the entire testing set for manual labeling via a **query function** (pool-based, batch-mode AL)

**Heuristic:** Query data in regions of feature space that are densely populated with testing data and sparsely populated with training data.
Methods: Active Learning (AL)

- $\hat{P}_{RF}(y|x)$ is the estimated RF probability
- $\rho(x', x)$ is the RF proximity measure

Proposed RF AL query functions; Richards et al. (2011)

AL1. Select testing data point ($x' \in \mathcal{U}$) that is most under-sampled by the training data ($\mathcal{L}$):

$$S_1(x') = \frac{P_{Test}(x')}{P_{Train}(x')} \approx \frac{\sum_{x \in \mathcal{U}} \rho(x', x)/N_{Test}}{\sum_{z \in \mathcal{L}} \rho(x', z)/N_{Train}}$$

(1)

AL2. Select testing data point that maximizes the total change in the RF probabilities over the testing data:

$$S_2(x') = \frac{\sum_{x \in \mathcal{U}} \rho(x', x)(1 - \max_y \hat{P}_{RF}(y|x))}{\sum_{z \in \mathcal{L}} \rho(x', z) + 1}$$

(2)
Experiment: OGLE+Hipparcos
Experiment

1542 OGLE+Hip sources (25 classes) randomly split into:
Training (black ▲)
Testing (red ×)

Selection function:

$$\Gamma \propto \log(P) \log(A)^{1/4}$$

We compare methods based on error rate on testing data
Experiment: Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active Learning</td>
<td></td>
</tr>
<tr>
<td>ST / CT</td>
<td></td>
</tr>
<tr>
<td>IW</td>
<td></td>
</tr>
<tr>
<td>&quot;Random&quot; AL</td>
<td></td>
</tr>
</tbody>
</table>

Note: AL routines evaluated only on non-selected testing data
Application: ASAS
We use AL to classify all 50124 sources in the ACVS catalog

Training set: 1542 well-understood stars (in 25 classes) from OGLE+Hipparcos

Perform 9 AL iterations of 50 sources each selected by sampling from $S$

Incorporate labeling “cost”:

$S(x) = S_2(x)(1 - C(x))$

11 users labeled sources. Use IEThresh crowd-sourcing of Donmez et al. (2009)
Performance metrics of classifier vs. AL iteration:
Results: ASAS

**AL classifications, compared to ACVS:**

<table>
<thead>
<tr>
<th>ACVS</th>
<th>AL-RF</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Mira</td>
<td>0.9</td>
</tr>
<tr>
<td>f. Classical Cepheid</td>
<td>0.6</td>
</tr>
<tr>
<td>g. Pop. II Cepheid</td>
<td>0.04</td>
</tr>
<tr>
<td>h. Multi. Mode Cepheid</td>
<td>0.05</td>
</tr>
<tr>
<td>j. Delta Scuti</td>
<td>0.01</td>
</tr>
<tr>
<td>k. Beta Cephei</td>
<td>0.01</td>
</tr>
<tr>
<td>q. Chem. Peculiar</td>
<td>0.02</td>
</tr>
<tr>
<td>w. Beta Persei</td>
<td>0.02</td>
</tr>
<tr>
<td>x. Beta Lyrae</td>
<td>0.01</td>
</tr>
<tr>
<td>y. W Ursae Maj.</td>
<td>0.01</td>
</tr>
</tbody>
</table>

**Time Series Classification**

J. Richards
A Related Issue: Noisification

“Noisification” and “Denoisification” approaches both dominate off-the-shelf classifier

See Long et al. (2011) in prep.
Some Current & Future Work

▶ Automated classification for time-domain surveys
  ▶ Anomaly detection See Sharmo Bhattacharyya’s poster
  ▶ Multi-band light curves: (1) leveraging multi-band observations for better period estimates, (2) extracting useful multi-band features
  ▶ Wavelet approaches for feature creation (Alexander Blocker & Debashis Mondal talks)

▶ Sample selection bias
  ▶ Transfer Learning (Ricardo Vilalta’s talk)
  ▶ How to start building training sets for Gaia / LSST?
  ▶ Resource allocation (AL) crucial for peta-scale data!
  ▶ Bayesian approaches: Bayesian experimental design, hierarchical models, ...
  ▶ How to deal with sample selection bias for parameter inference problems?


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”)
