Towards an Automated Classification of Transient Events in Synoptic Sky Surveys

S. G. Djorgovski, C. Donalek, A. Mahabal (Caltech)
B. Moghaddam, M. Turmon (JPL)
M. Graham, A. Drake, N. Sharma, Y. Chen (Caltech)

CIDU 2011, Mountain View, CA, October 2011
Astronomy in Time Domain

• Synoptic digital sky surveys – i.e., a panoramic cosmic cinematography – are now becoming the dominant data producers in astronomy
  – From Terascale to Petascale data streams

• A major new growth area of astrophysics
  – Driven by the new generation of large digital synoptic sky surveys, leading to LSST, SKA, etc.

• A broader significance for an automated, real-time knowledge discovery in massive data streams
Astronomy in the Time Domain

• Rich phenomenology, from the Solar system to cosmology and extreme relativistic physics
  – Touches essentially every field of astronomy
• For some phenomena, time domain information is a key to the physical understanding
• A qualitative change:
  Static $\Rightarrow$ Dynamic sky
  Sources $\Rightarrow$ Events
• Real-time discovery/reaction requirements pose new challenges for knowledge discovery

Synoptic, panoramic surveys $\Rightarrow$ event discovery
Rapid follow-up and multi-\(\lambda\) $\Rightarrow$ keys to understanding
The Catalina Real-Time Transient Survey (CRTS)

- Collaboration with UAz/LPL search for NEA/PHA asteroids
- 3 small telescopes up to 2,500 deg\(^2\)/night with 4 exposures/pointing, limiting mags \(\sim 19 – 21\), several tens of passes per year, total area coverage \(\sim 33,000 \text{ deg}^2\), time baselines from 10 min to years, \(\sim 7+\) years coverage
- Real time processing and event discovery and publication
- Open data policy: *all data are made public immediately*
- \(\sim 4,000\) unique transients so far, a number of discoveries made
Sample Light Curves

The plan is to produce light curves for every detected source in the survey (> $5 \times 10^8$ sources), make them publicly available, and mine that data set. Light curves are generated on demand for transient sources, blazars, etc.
Automated Classification of Transients

Vastly different physical phenomena, yet they look the same! Which ones are the most interesting and worthy of follow-up?

Rapid, automated transient classification is a critical need!
This is a Critical Problem
(and it will get a lot worse)

• Now: data streams of \(~ 0.1 \text{ TB} / \text{night}, \sim 10^2 \text{ transients} / \text{night}\) (CRTS, PTF, various SN surveys, microlensing, etc.)
  ♦ We are already in the regime where we cannot follow them all
  ♦ Spectroscopy is the key bottleneck now, and it will get worse

• Forthcoming on a time scale \sim 1 - 5 \text{ years}: 
  \sim 1 \text{ TB} / \text{night}, \sim 10^3 - 10^4 \text{ transients} / \text{night} 
  (PanSTARRS, Skymapper, VISTA, VST, SKA precursors...)

• Forthcoming in \sim 8 - 10 (?) \text{ years}: LSST, \sim 30 \text{ TB} / \text{night}, \sim 10^5 - 10^7 \text{ transients} / \text{night}, SKA

• So... which ones will you follow up?

• Follow-up resources will likely remain limited

\text{Transient classification is essential}
Event Classification is a Hard Problem

• Classification of transient events is essential for their astrophysical interpretation and uses
  – Must be done in real time and iterated dynamically

• Human classification is already unsustainable, and will not scale to the Petascale data streams

• This is hard:
  – Data are sparse and heterogeneous: feature vector approaches do not work; using Bayesian approach
  – Completeness vs. contamination
  – Follow-up resources are expensive and/or limited: only the most interesting events
  – Iterate classifications dynamically as new data come in

• Traditional DP pipelines do not capture a lot of the relevant contextual information, prior/expert knowledge, etc.
Towards the Automated Event Classification

- Incorporation of the contextual information (archival, and from the data themselves) is essential
- Automated prioritization of follow-up observations, given the available resources and their cost
- A dynamical, iterative system
Automated Detection of Artifacts

Automated classification and rejection of artifacts masquerading as transient events in the PQ survey pipeline, using a Multi-Layer Perceptron ANN

Lead: C. Donalek
Generating priors for various observables for different types of variables

Lead: A. Mahabal
Bayesian Networks Implementation

\[ X = \text{input measurements of individual kinds (e.g., mags, colors, etc.)} \]
\[ Y = \text{classes of events, } Y = 1, \ldots, k \]

Then:

\[ P(y = k \mid x) = P(x \mid y = k)P(k)/P(x) \propto \]

\[ \propto P(k)P(x \mid y = k) \approx P(k)\prod_{b=1}^{B} P(x_b \mid y = k) \]

Initial results using single-epoch color measurements:
Typical accuracy ~ 80%
Typical contamination ~ 15 – 20%

Expecting significant improvements when more observed features are used
A Hierarchical Approach to Classification

Different types of classifiers perform better for some event classes than for the others.

We use some astrophysically motivated major features to separate different groups of classes.

Proceeding down the classification hierarchy, every node uses those classifiers that work best for that particular task.
LC Feature Vectors

- Create a new dataset extracting features from CRTS light curves (CV, Blazar, RR Lyrae, SN, etc)
- Each set of light curve features become a row in the new dataset
- Apply feature vector methodology

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>amplitude</td>
<td>Half the difference between the maximum and the minimum magnitude</td>
</tr>
<tr>
<td>beyond1std</td>
<td>Percentage of points beyond one st. dev. from the weighted mean</td>
</tr>
<tr>
<td>flux_percentile_ratio_mid20</td>
<td>Ratio of flux percentiles (60th - 40th) over (95th - 5th)</td>
</tr>
<tr>
<td>flux_percentile_ratio_mid35</td>
<td>Ratio of flux percentiles (67.5th - 32.5th) over (95th - 5th)</td>
</tr>
<tr>
<td>flux_percentile_ratio_mid50</td>
<td>Ratio of flux percentiles (75th - 25th) over (95th - 5th)</td>
</tr>
<tr>
<td>flux_percentile_ratio_mid65</td>
<td>Ratio of flux percentiles (82.5th - 17.5th) over (95th - 5th)</td>
</tr>
<tr>
<td>flux_percentile_ratio_mid80</td>
<td>Ratio of flux percentiles (90th - 10th) over (95th - 5th)</td>
</tr>
<tr>
<td>linear_trend</td>
<td>Slope of a linear fit to the light curve fluxes</td>
</tr>
<tr>
<td>max_slope</td>
<td>Maximum absolute flux slope between two consecutive observations</td>
</tr>
<tr>
<td>median_absolute_deviation</td>
<td>Median discrepancy of the fluxes from the median flux</td>
</tr>
<tr>
<td>median_buffer_range_percentage</td>
<td>Percentage of fluxes within 20% of the amplitude from the median</td>
</tr>
<tr>
<td>pair_slope_trend</td>
<td>Percentage of all pairs of consecutive flux measurements that have positive slope</td>
</tr>
<tr>
<td>percent_amplitude</td>
<td>Largest percentage difference between either the max or min magnitude and the median</td>
</tr>
<tr>
<td>percent_difference_flux_percentile</td>
<td>Diff. between the 2nd &amp; 98th flux percentiles, converted to magnitude²</td>
</tr>
<tr>
<td>QSO</td>
<td>Quasar variability metric in Butler &amp; Bloom (2010)</td>
</tr>
<tr>
<td>non_QSO</td>
<td>Non-quasar variability metric in Butler &amp; Bloom (2010)</td>
</tr>
<tr>
<td>skew</td>
<td>Skew of the fluxes</td>
</tr>
<tr>
<td>small_kurtosis</td>
<td>Kurtosis of the fluxes, reliable down to a small number of epochs</td>
</tr>
<tr>
<td>std</td>
<td>Standard deviation of the fluxes</td>
</tr>
<tr>
<td>stetson_j</td>
<td>Welch-Stetson variability index Jb</td>
</tr>
<tr>
<td>stetson_k</td>
<td>Welch-Stetson variability index Kb</td>
</tr>
</tbody>
</table>

Richards et al. (2011)
Optimizing Feature Selection

Select a subset of features from the data matrix $X$ that best predict the data in classes $Y$ by sequentially selecting features until there is no improvement in prediction: using Decision Trees with a 10-fold cross validation.

<table>
<thead>
<tr>
<th></th>
<th>Completeness</th>
<th>Contamination</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blazar</td>
<td>83%</td>
<td>13%</td>
</tr>
<tr>
<td>CV</td>
<td>94%</td>
<td>6%</td>
</tr>
<tr>
<td>RR Lyrae</td>
<td>97%</td>
<td>4%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Completeness</th>
<th>Contamination</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blazar</td>
<td>81%</td>
<td>13%</td>
</tr>
<tr>
<td>CV</td>
<td>96%</td>
<td>5%</td>
</tr>
<tr>
<td>SN Ia</td>
<td>100%</td>
<td>&lt;1%</td>
</tr>
</tbody>
</table>

Amplitude beyond 1 std
flux_percentile_ratio_mid65
max_slope
qso
std
lomb-scargle

Linear_trend
Median_absolute_deviation
lomb-scargle

(N. Sharma)
2D Light Curve Priors

- For any pair of light curve measurements, compute the $\Delta t$ and $\Delta m$, make a 2D histogram
  - Note: $N$ independent measurements generate $N^2$ correlated data points
- Compare with the priors for different types of transients
- Repeat as more measurements are obtained, for an evolving, constantly improving classifier.

Lead: B. Moghaddam
Applying \( \Delta m \) vs. \( \Delta t \) Histograms

- Measure of a divergence between the unknown transient histogram and two prototype class histograms
**Δm vs. Δt Classifier Performance**

- Performance measured using Leave-one-out cross-validation (LOOCV)

<table>
<thead>
<tr>
<th></th>
<th>SN</th>
<th>CVBlazarRRMira</th>
</tr>
</thead>
<tbody>
<tr>
<td>SN</td>
<td>A0 = 96.5%</td>
<td>3.5%</td>
</tr>
<tr>
<td>CVBlazarRRMira</td>
<td>2.1%</td>
<td>A1 = 97.9%</td>
</tr>
</tbody>
</table>

- Optimize histogram parameters (binning, smoothing, Dirichlet prior parameters) using a genetic algorithm

<table>
<thead>
<tr>
<th></th>
<th>SN</th>
<th>CVBlazarRRMira</th>
</tr>
</thead>
<tbody>
<tr>
<td>SN</td>
<td>99.3%</td>
<td>0.7%</td>
</tr>
<tr>
<td>CVBlazarRRMira</td>
<td>1.5%</td>
<td>98.5%</td>
</tr>
</tbody>
</table>

- A modest, but a consistent improvement over the human expert selected parameters (Y. Chen)
New Machine Discovery Tools

Hod Lipman et al. (Cornell)

Initial experiments on using Eureqa for an automated classifier discovery (M. Graham et al.)
Eureqa: How It works
http://creativemachines.cornell.edu/eureqa

• Employs symbolic regression to determine best-fitting functional form to data – fits both form of equation and its parameters simultaneously
• Specify type of building block to be used in a fit: algebraic operators, analytical functions (trig, exp), constants, state variables
• Works from numerical partial derivatives of each pair of variables
• Employs a genetic-type algorithm to explore pd-metric space
• Produces a small set of final candidate analytical expressions on accuracy-parsimony Pareto front
Classifying Light Curves with Eureqa

• Light curves of two known stellar classes:

  ![Eclipsing binary (W U Ma)](image1)
  ![RR Lyrae](image2)

• Characterize with ~60 periodic/non-periodic features (Richards et al. 2011, Debosscher et al. 2007)

• Use Eureqa for binary classification: \textit{class 1} vs. \textit{class 2}

• Fit: \texttt{class = step[f(x_1, x_2, x_3, \ldots, x_{60})]}

Eureqa: Initial Results

- CRTS light curves for:
  - 222 W U Ma in GCVS
  - 417 RR Lyrae in SDSS Stripe 82

<table>
<thead>
<tr>
<th></th>
<th>W U Ma</th>
<th>RR Lyrae</th>
</tr>
</thead>
<tbody>
<tr>
<td>W U Ma</td>
<td>89.6%</td>
<td>10.4%</td>
</tr>
<tr>
<td>RR Lyrae</td>
<td>2.4%</td>
<td>97.6%</td>
</tr>
</tbody>
</table>

- Best fitting function:
  \[ \text{Class} = \text{step}[\cos(4.21x_{20} - x_{25}) - x_{14}] \]

- Involved features are:
  - \( x_{14} \): Ratio of (95\textsuperscript{th} – 5\textsuperscript{th}) flux percentile over median flux
  - \( x_{20} \): Kurtosis (reliable for small sample sizes)
  - \( x_{25} \): False-peak probability of top period in L-S periodogram

- Very promising!

Lead: M. Graham
Contextual Information is Essential

- **Visual context** contains valuable information about the reality and classification of transients.
- So does the **temporal context**, from the archival light curves.
- And the **multi-λ context**.
- Initial detection data contain little information about the transient: $\alpha$, $\delta$, $m$, $\Delta m$, $(t_c)$. **Almost all of the initial information is archival or contextual**; follow-up data trickle in slowly, if at all.
Harvesting Human Pattern Recognition

Human-annotated images (via SkyDiscovery.org)

- Semantic descriptors
- Machine processing
- Evolving novel algorithms

Challenges: Optimizing for different levels of user expertise; optimal input averaging; encoding contextual information; etc.
Work in Progress: Fusion Module

Colors and light curve information can be combined in one network. This "fusion module" combines the probabilistic results from each constituent classifier.

Exploring a variety of techniques for optimal classification fusion:

- Markov Logic Networks
- Diffusion Maps
- Multi-Arm Bandit
- Sleeping Expert...

\( P_{\text{class}} \)
Automating the Optimal Follow-Up

For the *potentially most interesting events*, what type of follow-up observations $a x$ has the greatest potential to discriminate among the competing event classes $y$?

Request the optimal follow-up observations from the available assets that maximize the entropy drop:

$$H[p(y | x_+, x_0)] = - \sum_{y, x_+} p(y, x_+ | x_0) \log p(y | x_+, x_0)$$
Summary

• Time domain astronomy is a vibrant new research frontier

• A key challenge: a real-time, automated, iterated transient event classification
  – Sparse and heterogeneous data, real time, dynamically iterated, resource-limited

• Promising results so far
  – Mostly Bayesian approaches, some light curve feature vector classifiers, AI/ML
  – A hierarchical approach to event classification
  – Next: an automated decision making for optimal follow-up observations
  – Harvesting human pattern recognition skills and expertise using citizen science / crowdsourcing

• A broader relevance for a real-time mining of massive data streams
Announcing a Workshop:

Digging Faster and Deeper:

Algorithms for Computationally Limited Problems in Time-Domain Astronomy

Caltech, Pasadena, CA, Dec. 12-13, 2011

http://www.astro.caltech.edu/digging