Classification with Sparse Timeseries

Ashish Mahabal
aam@astro.caltech.edu
7 Sep 2011
HEAD 2011, Newport, RI
Collaborators

• Caltech
  – George Djorgovski
  – Ciro Donalek
  – Andrew Drake
  – Matthew Graham
  – Roy Williams

• JPL
  – Baback Moghaddam
  – Mike Turmon

Plus at various other institutes all over, but especially in US, India and Italy
Semantic Tree of Astronomical Variables and Transients

Asteroids
  - Rotation
  - Eclipse
    - Asteroid occultation
    - Planetary transits

Stars
  - Eclipse
    - Microlensing
  - Rotation
  - Eruptive
    - Cataclysmic
  - Pulsation
    - Secular

SN Subtypes

Mostly optical viewpoint

AGN Subtypes

Credit: L. Eyre & N. Mowlavi (10/2007)
Sample Light Curves

Blazar PKS0823+033

CV 111545+425822

Variables and transients – the distinction is one of perception, and your aims

9/7/11

Ashish Mahabal
Towards Automated Event Classification

A necessity for large synoptic surveys

Event parameters: $m_1(t), m_2(t), \ldots$, $\alpha$, $\delta$, $\mu$, ... image shape...

Event Classification Engine

P(SN Ia) = ...
P(SN II) = ...
P(AGN) = ...
P(CV) = ...
P(dM) = ...

... etc.

Classification probabilities (evolving, iterated)

With M Turmon and B Moghaddam, JPL
Making optimal use of sparse data sets

• sparse light curves
  • analysis of different types
• few colors/sparse SEDs
• any contextual information
• priors for different kinds of objects

Holistic approach
Catalina Sky Survey(s):

CRTS uses the data from all three Catalina NEO surveys, with a coverage of up to 2,500 deg² / night, and the total area coverage of ~ 32,000 deg²

<table>
<thead>
<tr>
<th>Survey region (deg)</th>
<th>+/- 5 deg ecliptic</th>
<th>-25 &lt; Dec &lt; +70</th>
<th>-80 &lt; Dec &lt; -25</th>
</tr>
</thead>
<tbody>
<tr>
<td>Field of View (square deg)</td>
<td>1.2</td>
<td>8.1</td>
<td>4.2</td>
</tr>
<tr>
<td>Mag limit (V)</td>
<td>21.5</td>
<td>19.5</td>
<td>19.0</td>
</tr>
</tbody>
</table>

We are processing the Catalina data streams in real time to look for astrophysical transients

9/7/11
CSS coverage
Follow-Up Observations:

- **Photometry** *(P60, NMSU, DAO, HTN, India, Mexico, etc.)*
- **Spectroscopy** *(Gemini N+S, Keck, P200, SMARTS, IGO, MDM)*

CSS090421:174806+340401  A blazar, also monitored at OVRO in radio
CRTS Event Detections

Distinct Events Detection Statistics as of 5 Jun 2011 UT:

<table>
<thead>
<tr>
<th>Tel</th>
<th>All OTs</th>
<th>SNe</th>
<th>CVs</th>
<th>Blazars</th>
<th>Ast/Flares</th>
<th>CV/SN</th>
<th>AGN</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSS</td>
<td>2033</td>
<td>596</td>
<td>501</td>
<td>113</td>
<td>184</td>
<td>275</td>
<td>229</td>
<td>195</td>
</tr>
<tr>
<td>MLS</td>
<td>1560</td>
<td>183</td>
<td>38</td>
<td>12</td>
<td>122</td>
<td>374</td>
<td>744</td>
<td>214</td>
</tr>
<tr>
<td>SSS</td>
<td>227</td>
<td>24</td>
<td>93</td>
<td>7</td>
<td>5</td>
<td>43</td>
<td>16</td>
<td>42</td>
</tr>
<tr>
<td>Total</td>
<td>3820</td>
<td>803</td>
<td>632</td>
<td>132</td>
<td>311</td>
<td>692</td>
<td>989</td>
<td>451</td>
</tr>
</tbody>
</table>

- Threshold set deliberately very high – only the most dramatic transients are pulled out in the real time
- About 1 strong transient per $10^6$ source detections
- The rate of significant transients/variables is at least an order of magnitude higher
- Many events are re-detected repeatedly (not counted above)

9/7/11

Ashish Mahabal
SSS and CSS transients
The output is BN class which is fed to skyalert as an annotation to the original event.

CV/SN classification ~80% with single epoch.
Naïve Bayes

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

- \( x \): feature vector of event parameters
- \( y \): object class that gives rise to \( x \) (1<\( y \)<\( k \))
- Certain features of \( x \) known: (position, flux)
- Others will be unknown: (color, delta-mag)
- Assumption: based on \( y \), \( x \) is decomposable into \( B \) distinct independent classes (labeled \( x_b \))
- This helps with the curse of dimensionality
- Also allows us to deal with missing values
The importance of context

Which galaxy does a supernova belong to?

The need to see the big picture
Characterization Vs. Classification

• Early focus on the extraction and dissemination of time series
• Characterizations is important
  – $\frac{dm}{dt}$
  – change of direction per unit time
  – change in periodicities (e.g., wavelet or fourier decomposition);
  – variation in $\frac{dm}{dt}$
  – acceleration in $\frac{dm}{dt}$

Most SNe will not become fainter and then brighten up
Aspects of \( \frac{dm}{dt} \) processing

- \( \frac{dm}{dt} \) features capture sparse or irregular LCs
- The features, and thus the underlying density models, are invariant to absolute magnitude and time shifts
- Features & densities allow bound-only flux observations
  - Under poor seeing, we obtain only bounds like \( m > 18 \)
By taking subsections of $\Delta t/\Delta m$ space determine which area is characteristic for which kind of variable.

9/7/11
Ashish Mahabal
Decision Tree decomposes this multi-class classifier into a series of binary discrimination tasks.

This specific DT follows the stratification that seems natural to astronomers.

All nodes shown were implemented via dm/dt histogram binary classifiers.

9/7/11
Ashish Mahabal
For a specific dm/dt region

> prep_Blazar.CV
> RUNME_prototype_classifier

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Blazar accuracy</td>
<td>83.33 %</td>
</tr>
<tr>
<td>CV accuracy</td>
<td>58.85 %</td>
</tr>
<tr>
<td>Total accuracy</td>
<td>64.31 %</td>
</tr>
<tr>
<td>Average accuracy</td>
<td>71.09 %</td>
</tr>
</tbody>
</table>

9/7/11

Ashish Mahabal
Using GAs to determine intervals

• *dmagbins*
  – generate array elements (intervals) based on a normal distribution around 0
  – remove negative elements in array
  – sort in ascending order
  – bin marks determined by cumulative sum of array elements
  – reflect over 0 to build symmetric *dmagbins*

• *dtbins*
  – 1-day interval for first 25-35 days (chosen uniformly at random)
  – 5-day interval for next 30 days
  – 10-day interval for next 60 days
  – 20-day interval for next 240 days
  – 365-day interval until end

• $\sigma_{dm}$: chosen uniformly at random from [0.2, 1.5]
• $\sigma_{dt}$: chosen uniformly at random from [0.2, 1.5]
• *SymDir*: 0 or 1
• *Alpha*: chosen log-uniformly at random from $[10^{-3}, 10^{2}]$
dm/dt bins as selected by GA
Automating the Optimal Follow-Up

What type of follow-up data has the greatest potential to discriminate among the competing models (event classes)?

Request follow-up observations from the optimal available facility.

Collaboration with B. Moghaddam, M. Turmon (JPL)

Ashish Mahabal
Event Publishing / Dissemination
skyalert.org

PI: R. Williams

• Real time:
  – VOEvents, Twitter, iApp (thousands of events)
  – Also on SkyAlert.org, feeds to the WWT, GoogleSky

• Next day: annotated tables on the CRTS website
Transient classification mantra

• Obtain a couple of epochs in one or more filters
• Assigns probabilities for different classes
• Choose observations (filters, wavelengths) for best discrimination
• Feed the new observations back in
• Revise probabilities, choose observations, ...
• Based on confirmed class revise priors

Bayesian network, dm/dt processing, (DAME, VOStat, VO), Skyalert

Ashish Mahabal
INPUT PARAMETERS

TRIGGER RULES

EXTERNAL & CONTEXTUAL INFORMATION

Expert 1: GPR

Expert 2: BN

Class

Expert 3: NN

\( P_1 \)

\( P_2 \)

\( P_3 \)

FUSION MODULE

CLASSIFIER FRAMEWORK

9/7/11

Ashish Mahabal
Bayesian Network/fusion modules are no Cartesian theatre

- Different parameters, methods are separate (though perhaps not independent) probes

(non-)Cartesian theatre
One observation can drive the direction given the large number of possible candidates
Not much scope for error