January 7 - March 2, 1610
Real-Time Knowledge Extraction from Massive Time-Series Datastreams

Josh Bloom
Astronomy Department

jbloom@astro.berkeley.edu
Extragalactic Transient Universe: Explosive Systems

Type Ia
NS + NS Mergers
IMBH + WD Collision
NS + RSG Collision
Pair Production Supernovae

Days Since Explosion

Brightness

$M_H$
“Bad” News: Discoveries Swamp Followup Resources

Large Synoptic Survey Telescope (LSST):
1 Gb every 2 seconds

$10^6$ supernovae/yr
$10^5$ eclipsing systems
$10^7$ asteroids...

light curves of 800 million sources every 3 days
Transients Classification Project

Berkeley Astronomy:
Dan Starr, Dovi Poznanski, Maxime Rischard, Nat Butler, Chris Klein, Rachel Kennedy, Justin Huggins, Adam Morgan, Adam Miller, JSB

San Francisco State University:
John M. Brewer

Berkeley Statistics:
Noureddine El Karoui, John Rice

Berkeley CS:
Martin Wainwright, Masoud Nikravesh

Lawrence Berkeley Lab:
Peter Nugent, Horst Simon

Los Alamos Nat. Lab. / UC Santa Cruz:
Damian Eads
Goal: Autonomous creation of new knowledge, that itself spurs further resource allocation & inquiry

- Generate **probabilistic statements** about the nature of events (ie. classification)
- Provide push/pull **access** to current & past events
- (bootstrap) Learning from feedback
- Operate at sufficient & **scalable** rates
Considerable Complications with Time Series Data

- noisy, irregularly sampled
- spurious data
- telltale signature event may not have happened yet

class: microlensing
2D image classification: Machine-Learning with human input

>1000:1 rejection of bogus candidates (prelim. cuts + machine learning)
most subtractions are bogus...

...but a long tail of astrophysical goodness

275,000 very likely real

10M PTF subtractions (1 month of data)
Major Challenge:

how do we use *domain knowledge* & *known ("labelled") instances* to create a classifier?

*traditional fitting, machine learning, ...*
## Machine-Learning Approach to Classification

<table>
<thead>
<tr>
<th>Data</th>
<th>Utility for Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Time Series</strong></td>
<td>• comparison to previously observed sources, &amp; theoretical/numerical models</td>
</tr>
<tr>
<td>(e.g. color, brightness change, etc.)</td>
<td>• historical images: extend time baseline</td>
</tr>
<tr>
<td><strong>Context</strong></td>
<td>situational awareness: expectations of different classes</td>
</tr>
<tr>
<td>(e.g. sky location, nearest galaxy type)</td>
<td></td>
</tr>
</tbody>
</table>

- **Less data regime**
- **More data regime**
- **Context**
- **Time-series**
Feature Extraction: Homogenizing Heterogenous Data

“Features”: real-number metrics that describe the time-domain characteristics & context of a source.

**variability metrics:**
e.g. Stetson indices, $\chi^2$/dof (constant hypothesis)

**periodic metrics:**
e.g. dominant frequencies in Lomb-Scargle, phase offsets between periods

**shape analysis**
e.g. skewness, kurtosis, Gaussianity

**context metrics**
e.g. distance to nearest galaxy, type of nearest galaxy, location in the ecliptic plane
Fig. 9. The classification based on $R^2_{11}$ obtained from the FD method. In the electronic version, coloured filled circles and upper triangles denote the stars from LMC and SMC respectively.

Mira variables, there is some overlap in the regions dominated by RR Lyraes and Cepheids. In the next step, we choose only the samples of RR Lyraes (RRab & RRc) and Cepheids (FU & FO) that could not be separated well by the use of PCA on the whole data set. We now run PCA on 10,643 light curves (Data set IA + IB + IIA + IIB + IIC + III) of RR Lyraes and Cepheids. The result of PCA on this 10643 × 100 array is shown in Fig. 12. It may be noticed that PC1 is able to separate FU Cepheids and RRab stars to a large extent while there is some overlap between RRc and FO Cepheids in a narrow period range (0.25-0.5 d). We hope to return to this degeneracy problem in a subsequent study in which we also intend to increase the sample by adding more classes of variables.

5. Conclusions

Fourier decomposition is a trusted and much applied technique for analyzing the behaviour of light curves of periodic variable stars. It is well suited for studying individual light curves as the Fourier parameters can be easily determined. However, when the purpose is to tag a large number of stars for their variable class using photometric data from large surveys, the technique becomes slow and cumbersome and each light curve has to be fitted individually and then analyzed. Same is true if the aim is to look for resonances in the light curves in an automated way for a large class of pulsators. It is, therefore, desirable to look for methods that are reliable, automated and unsupervised and can be applied to the available light curve data directly.

Some attempts have been made in the recent past to use the well known PCA for the light curve analysis, but the major drawback of these studies was that they required the calculation of the Fourier parameters which then went as input to the PCA. This meant that the PCA, which was supposed to replace the Fourier decomposition, in fact relied on it. Also for precise and accurate determination of Fourier parameters, the light curve should have good phase coverage and less noisy data points so that the fit to the light curve is good enough to rely on its parameters. But this is not the case for each and every light curve data generated from the automated surveys. Sometimes there are gaps and/or outliers in the data. The fitting of such a light curve will give a wrong estimation of the Fourier parameters.

In this paper we have used the original light curve data for computation of the principal components. It involves four simple steps.

**Fig. 10.** The classification based on $R^2_{11}$ obtained from the FD method.

**Fig. 11.** The classification based on PC1 obtained from PCA of 100 interpolated magnitudes for the phase from 0 to 1 in steps of 0.01.

Deb & Singh+09
Transient Taxonomy is a Mess

Phenomenologically & physically based taxonomy

- hybrid topology → complicates machine learning
- incomplete & inaccurate

Can the learning process itself, based purely on what is observed, reveal new physical connections between phenomena?
Machine-Learned Classification
1. Parallelize the Learning Phase of Machine Learning

**Problem:**
frameworks like Weka ([http://www.cs.waikato.ac.nz/ml/weka/](http://www.cs.waikato.ac.nz/ml/weka/)) are not natively parallel. We will need to burst out training requests on specific time/observation vectors & classify quickly with the results

**Solution:**
- build a parallel platform for weka (GridWeka, Weka-parallel etc. are out of date & probably not not elegant)

[http://userweb.port.ac.uk/~khusainr/weka](http://userweb.port.ac.uk/~khusainr/weka)
1. Parallelize the Learning Phase of Machine Learning

**Problem:**
we have errors on our data (both training sets and instances) & we don't know how to deal with them

**Sledgehammer Solution:**
use a parallel platform to generate distribution of trained models & apply to distribution of instance-based sets
1. Parallelize the Learning Phase of Machine Learning

- Flux
- 68% confidence interval
- "fastest rise" feature extractor
- Time
- 0.2
- 0.5
- 0.32
2. Build a General Crowdsourcing Platform (GroupThink2.0)
- production scale site (GoogleAppEngine or elsewhere), allowing interconnection of projects
2. Build a General Crowdsourcing Platform (GroupThink2.0)
- build innovative analytics plugins for projects;
- could require grid/cloud-based analysis for on-the-fly results
3. Parallelized Genetic Programming for Feature Discovery

Instead of handcoding “features” for ML, using GP (in parallelized environment) to *discover features* which give the best classification.

![Cepheid vs. RR Lyrae graphs](Image)
4. Parallelized Visual Exploration Tool

allow the armchair astronomer to ask complex questions of the databases & visualize and interact with the results (100M+ rows)
4. Parallelized Visual Exploration Tool

allow the armchair astronomer to ask complex questions of the databases & visualize and interact with the results

- parallel database calls with embedded custom code (e.g. Hadoop SQL “hive”)

![Google Finance](image-url)
Resources

1. dotastro.org

2. Harvard TimeSeries Center:
   http://timemachine.iic.harvard.edu/
