Commentary on Richards, "Time-Series Classification in the Synoptic Surveys Era"

Chad M. Schafer
www.stat.cmu.edu/~cschafer
Department of Statistics
Carnegie Mellon University
June 2011
Acknowledgements

Stefan Etchegaray

Peter Freeman

Ashish Mahabal
The Challenge

From the *LSST Science Book* (http://www.lsst.org/lsst/scibook):

“The main challenges ahead of massive time-domain surveys are timely recognition of interesting transients in the torrent of imaging data, and maximizing the utility of the follow-up observations.”
The Challenge

From the *LSST Science Book* (http://www.lsst.org/lsst/scibook):

“The very first observations of a transient may not reveal its class right away, and follow-up photometric observations will be required for a very large number of objects. Here too there will be a choice between different bands, available apertures, and sites. For example, follow-up with a specific cadence may be necessary for a suspected eclipsing binary, but with a very different cadence for a suspected nova. Follow-up resource prioritization can be done by choosing a set-up that reduces the classification uncertainty most.”
A Sequence of Data Challenges?

These are (significant) statistical/ML challenges

Provide concise problem descriptions, with accompanying data

A role for the *Informatics and Statistics Science Collaboration (ISSC)*?
Example: Blazar versus CV Classification

Blazar – quasars viewed from a particular angle

Cataclysmic Variable (CV) – Binary system, accretion leads to irregular outbursts

Both characterized by irregular variability in measured light curves

Data from Catalina Sky Survey, via Ashish Mahabal
CV
ID = CV103160010124100073

Apparent Magnitude

MJD − 53329.240005
A road map for light curve classification:

Light Curves

Features

Classes

Learning

Random Forest

Structured Classification

Prediction

known

unknown

Cepheids

IIp SN

Mira

QSO

RRL

Ia SN
Consider features of the **Identifying Data** – Those data, observed only on the training set, that permits labelling
SDSS DR8 Sky Server – http://skyserver.sdss3.org/dr8/
**Summary data for: SDSS J103351.42+605107.6**

**Position Data (How do I find it?)**

<table>
<thead>
<tr>
<th>Object ID (objID):</th>
<th>Right ascension (ra):</th>
<th>Declination (dec):</th>
</tr>
</thead>
<tbody>
<tr>
<td>1237653619081478229</td>
<td>158.46426112</td>
<td>60.65211188</td>
</tr>
</tbody>
</table>

**Image Data**  
(What does it look like?)

*Preview image (click to go to Navigate tool)*

- Object Type (type): STAR

**Spectrum Data**  
(What does its spectrum look like?)

*Preview spectrum (click for a larger version)*

- Spectral classification (Class): QSO

**Magnitudes:**

<table>
<thead>
<tr>
<th>Type</th>
<th>Magnitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ultraviolet (u)</td>
<td>19.32 ± 0.03</td>
</tr>
<tr>
<td>Green (g)</td>
<td>18.89 ± 0.01</td>
</tr>
<tr>
<td>Red (r)</td>
<td>18.47 ± 0.01</td>
</tr>
<tr>
<td>Infrared - 7600 Å (i)</td>
<td>18.19 ± 0.01</td>
</tr>
<tr>
<td>Infrared - 9100 Å (z)</td>
<td>17.93 ± 0.02</td>
</tr>
</tbody>
</table>

**Redshift Data:**

- Redshift (z): 1.408706

---

CV emission spectrum – via SDSS
Survey: sdss Program: legacy Target: QSO_SKIRT SERENDIP_BLUE
RA=158.46425, Dec=80.65204, Plate=560, Fiber=442, MJD=52298
z=1.40871±0.00069 Class=QSO BROADLINE
No warnings.

Blazar emission spectrum – via SDSS
Some feature of identifying data, call it $g(x_I)$, is great at separating objects into classes, and $g(x_I)$ can be predicted using observed data. Transformation can be fruitful. Even more idealistic: Sequential observations admit improved estimation of $g(x_I)$, with accurate quantification of uncertainty.
Active Learning

The motivation for using active learning is clear, this is another very challenging problem

Co-Training versus Joey’s Ideas: Label the “easy” points, or the “hard” points?

Again, the reliance upon physical understanding may be critical
Summary

Discussion of Time-Domain Classification:

The difficulty of the problem

Accepting the need to do probabilistic classification

The role of physical information