Towards a Real-time Transient Classification Engine

J. S. Bloom[1,□]**, D. L. Starr**[1,2]**, N. R. Butler**[1]**, P. Nugent**[1,3]**, M. Rischard**[1]**, D. Eads**[4]**,** and **D. Poznanski**[1]

[1] Astronomy Department, University of California, Berkeley, Berkeley, CA 94709 USA
[2] Las Cumbres Global Telescope Network, 6740 Cortona Dr. Santa Barbara, CA 93117 USA
[3] Lawrence Berkeley National Laboratory MS 50F-1650 1 Cyclotron Road Berkeley, CA, 94720 USA
[4] University of California, Santa Cruz 1156 High Street Santa Cruz, CA 95064 USA

Temporal sampling does more than add another axis to the vector of observables. Instead, under the recognition that how objects change (and move) in time speaks directly to the physics underlying astronomical phenomena, next-generation wide-field synoptic surveys are poised to revolutionize our understanding of just about anything that goes bump in the night (which is just about everything at some level). Still, even the most ambitious surveys will require targeted spectroscopic follow-up to fill in the physical details of newly discovered transients. We are now building a new system intended to ingest and classify transient phenomena in near real-time from high-throughput imaging data streams. Described herein, the *Transient Classification Project* at Berkeley will be making use of classification techniques operating on "features" extracted from time series and contextual (static) information. We also highlight the need for a community adoption of a standard representation of astronomical time series data (ie. "VOTimeseries").

1 Introduction

Classification and knowledge extraction from large imaging surveys of the static sky is a maturing endeavor. Star-galaxy classification and photometric redshift estimates are well posed problems where knowledge of the underlying physics allows for robust estimates of uncertainty. Automated classification is critical for source demographics, especially informing large statistical questions of the data. More recently, the autonomous classification with time domain surveys have had some notable successes, particularly with microlensing surveys 12 and supernova searches 4, 5, 9. Indeed, optimizing reductions, survey cadence strategies, and software algorithms in the search of a specific class of phenomena has been the most straightforward use of synoptic imaging (see Bailey, this workshop).While interested primarily in supernovae, the deep lens survey (2; see Becker, this workshop) is one of the few large-field synoptic surveys that attempted to provide multi-class inferences in real-time. Still, those classifications were rather broadbrushed in physical scope ("supernova", "fast moving", "slow moving", "variable star", "unknown (stationary)").With the advent of large-scalemulticolor imaging surveys to appreciable depths (e.g., DES, Pan-STARRS1, Skymapper, & LSST), the need for a realtime and general classification scheme of astronomical transient is particularly pressing.

The challenges in time-domain classificationmay be subdivided into "discovery" and "inference." Discovery of a variable or a moving source requires at least two images with the same filter system. The characteristic separation in time will dictate the types of sources that are seen to vary: taken too close in time, moving solar system sources do not change their apparent angular position enough to recognize change; taken too distant in time, slowly moving solar system sources would be single apparition point sources, confused for extra-solar events. Without the benefit of filtering techniques with several images of a static sky, transient discovery on a few images is particular prone to cosmetic defects in the imaging arrays, cosmic rays, and near-field (nonastrophysical) interlopers[1]. Transient discovery can be performed either in "catalog space" (noting significant changes of source brightness in two epochs) or with source extraction

in image differences 1. The former is generally less computationally intensive but more susceptible to error due to variable seeing and imaging array defects. Image differencing is generally more robust in crowded fields and where transients are embedded in galaxy light.

Once a transient source is discovered, we wish to surmise the nature of the source. Major features of a general classification scheme are identified:

– The inferences about the physical nature of the source (and the source variability) should make full use of prior knowledge about transients without coercing every new transient into a predefined set of classes. Different users should be allowed to tune their priors as the science requires.

– The inferences will necessarily be probabilistic in nature and should evolve in time as more observations are obtained.

– The classifications should be as near real-time as possible to allow appropriate follow-up.

– The classifications should allow feedback fromend users and adapt the classification algorithms accordingly.

We are now building a framework that will be capable of classifying transient sources fromtime-domain surveyswith these features. Similar work in a machine-learning context has been reported elsewhere (e.g., 13, 14 and Mahabal, this workshop; Bailey, this workshop). There are several components to this "Transient Classification Project" (TCP), described herein. We aim to have a working system in place by the time that the Palomar Transients Factory (PTF; 8) comes on-line in Fall 2008.

2 Data Ingest

The starting point of transient classification, from the perspective of the TCP, is a stream of metadata describing the individual detections (either from image differences or catalog detections). We coerce this metadata stream to an internal data model using a custom translation client written for each survey. Since we have been developing the TCP using, primarily, the public data from the SDSS-II stripe 82 survey 7, we found it necessary to recalibrate the photometry and astrometry from the raw detection files. Objects from all surveys are ingested into a relational database with the object positions indexed using the hierarchical triangle mesh (HTM; 10) at depth of 14 and 25 to allow for fast and accurate searching.

Object detections need to be associated with astrophysical sources. Unlike with static or after-the-fact time-domain surveys where a filtered deep sky image may be used as the fiducial "true" representation of source brightnesses and source positions, a real-time time-domain survey necessarily must associate each object detection with an astrophysical source. We create sources on-the-fly using a probabilistic framework that

asks the question whether a new object belongs to an existing source or demands the creation of a new source. For a new object with detected position ($O_\alpha \pm \sigma_{O,\alpha}, O_\delta \pm \sigma_{O,\delta}$) we can find a set of possible associated sources $\{S\}$ by searching in the source catalog for sources with positions with angular distance[2] d of O; typically we use $d_0 \approx 10$. For each source S_i in $\{S\}$ we compute the logarithm of the odds ratio $\log o_i$ comparing the hypotheses that the object belongs to that source or should be a separate source. Under the assumption of Gausianity with $\bar{S}_{\alpha,\delta}$ equal to the weighted mean position of O and S_i Since we expect $-2 \ln o_i$ to be distributed as χ^2 with 2 degrees of freedom we associate the object with the S_i meeting some predetermined probability threshold. As the number of objects associated with a source grows, the number of lower probability associations should also grow; through a simulation, we have found the correct number of sources are created if we change the probability threshold in accordance with the number of sources already associated with that object.

3 Source Classification

3.1 Representation

The creation or modification of a source triggers a series of steps that will lead to an updated statement about the nature of that source. In an effort to modularize the software tasks, and prepare for the possibility of distributing the computational tasks (see 4), we have build a portable (XML-based) source container (which we called "VOSource"), consisting of rudimentary source position, results from survey queries (such as NED), and the time series photometry associated with the source. The time series (which we call "VOTimeseries" is marked up as a VOTable[3], similar to the way in which time series data are represented in the VizieR catalogs. An example of a "VOSource" container is:

3.2 Feature Extraction:Mixing the Time-domain with Context

The time series of a source is anything but standard— data are irregularly sampled, noisy, sometimes spurious, andmay include detections as well as non-detections. In light of this, we seek to homogenize the time series data by extracting "features." A feature is a real-number line mapping, often involving basic statistical metrics on the time series (such as a chi-square per degree of freedom or skewness). We are developing a custom *Python* codebase which is capable of ingesting a VOTimeseries and returning the features of that time series. For sources with too few observations appropriate for a given feature (such as "largest significant peak in a Lomb-Scarle periodogram"), the results from that specific feature extraction is reported as undefined. One of the benefits of mapping heterogeneous information to a series of real-number lines is that the nature of that information is abstracted from operations performed on those feature vectors farther downstream. To this end, (static) context related to a transient source (e.g., the location with respect to the supergalactic plane, distance to nearest cataloged galaxy, redshift of that galaxy, etc.) can play a powerful discriminating role. For example, a new point source which is discovered close to the ecliptic plane and a region of significant Galactic extinction is much more likely to be a slow moving solar system object than a distant supernova.

3.3 Rapid Identification and Adaptation Using Machine Learning

The TCP will make use of prior knowledge of time series and contextual information for each known class of transient. To do so, we are assembling a large labeled training set of real-world examples of known classes. To create set of prior feature distributions, these sources will be degraded and sampled with cadences and sensitivities typical of the survey. A new source and its subsequently derived feature vector can be compared directly to the priors. We do not yet know what family of classifiers will prove the most robust (see Mahabal, this workshop, for an extensive discussion of current techniques). However, we are exploring the use of pairwise Naive Bayesian voting techniques as a fast approach capable of yielding probabilistic statements about the nature of new transients. Online learning algorithms 3 (such as "shifting experts" 6) and ensemble algorithms (such as "boosting" 11) should be particularly applicable, since these classes of algorithms allow quick updates of the classification inferences without needing to re-analyze all the available data. When a new event arrives that cannot be characterized by existing classes, methods for identifying the anomaly and incorporating it into a new class are being considered.

4 Future Steps

The TCP is a work in progress but the basic architectural decisions have now been put in place. Aside from the development and testing of the machine learning techniques, there are several other elements we hope the implement in the upcoming year:

- **Multi-survey Footprint Server**. Non-detections of a transient source provide valuable constraints for classification. A footprint server, providing upper-limits for a given position, time and filter, will be therefore crucial to the classification algorithms.
- **Distribution**.We plan to make full use of the VOEvent- Net architecture to distribute newly-classified transients to TCP clients (using a variety of push/pullmechanisms)
- **Feedback Mechanisms** We will require a formalized conduit for end users (on the receiving end of probabilistic classifications) to feed back in to the system the outcome of transient followup.
- **Massively Distributed Computing** We are scoping the use of the BOINC architecture5 to create a [TCP@Home](#) environment, where individual users will provide spare CPU cycles to crunch feature extraction methods and run classification algorithms.

While especially suited for the PTF, this classification engine is being built to not only allow several surveys streams to flow through the system but allow the information extracted in each stream to inform the classifications derived from other surveys. Since the implementation of the feature extraction and classification algorithms is atomized, we expect TCP to scale well to the data rates advertised for LSST.

References

Bailey, S., Aragon, C., Romano, R., et al.: 2007, ApJ , 665, 1246 Becker, A. C., Wittman, D. M., Boeshaar, P. C., et al.: 2004, ApJ , 611, 418

Blum, A.: 1998, In *Developments from a June 1996 seminar on Online algorithms*, Springer-Verlag, London, UK, ISBN 3-540-64917-4, 306

Copin, Y., Blanc, N., Bongard, S., et al.: 2006, New Astronomy Review, 50, 436

Frieman, J. A., Bassett, B., Becker, A., et al.: 2007, The Sloan Digital Sky Survey-II Supernova Survey: Technical Summary, astro-ph/0708.2749

Herbster, M. and Warmuth, M. K.: 1998, Mach. Learn., 32, 2, 151

Ivezi´c, ˇZ., Smith, J. A., Miknaitis, G., et al.: 2007, AJ , 134, 973

Kulkarni, S. R.: 2007, Palomar Transient Factory, http://www.mpa-garching.mpg.de/ grb07/Presentations/Kulkarni.pdf.

Miknaitis, G., Pignata, G., Rest, A., et al.: 2007, ApJ , 666, 674

O'Mullane, W., Banday, A. J., G´orski, K. M., et al.: 2001, In A. J. Banday, S. Zaroubi, and M. Bartelmann, editors, *Mining the Sky*, 638

Schapire, R. E. and Singer, Y.: 1999, Mach. Learn., 37, 3, 297

Udalski, A.: 2003, Acta Astronomica, 53, 291

Vestrand, W. T., Theiler, J., and Wozniak, P. R.: 2004, AN, 325, 477

Wozniak, P., Borozdin, K. N., Galassi, M. C., et al.: 2002, In *Virtual Observatories* A. S. Szalay, editor. Proceedings of the SPIE, 4846, 147