New Horizons in Time-Domain Astronomy
Proceedings IAU Symposium No. 285, 2011
R.E.M. Griffin, R.J. Hanisch & R. Seaman, eds.

doi:10.1017/S1743921312001056

Real-Time Classification of Transient Events in Synoptic Sky Surveys

Ashish A. Mahabal[1], C. Donalek[1], S. G. Djorgovski[1,2], A. J. Drake[1], M. J. Graham[1], R. Williams[1], Y. Chen[1], B. Moghaddam[3], and M. Turmon[3]

[1]California Institute of Technology, Pasadena, CA 91125, USA
email: aam@astro.caltech.edu

[2]Distinguished visiting professor, King Abdulaziz Univ., Jeddah, Saudi Arabia.

[3]Jet Propulsion Laboratory, Pasadena, CA 91109, USA

Abstract. An automated rapid classification of the transient events detected in modern synoptic sky surveys is essential for their scientific utility and effective follow-up when resources are scarce. This problem will grow by orders of magnitude with the next generation of surveys. We are exploring a variety of novel automated classification techniques, mostly Bayesian, to respond to those challenges, using the ongoing CRTS sky survey as a testbed. We describe briefly some of the methods used.

The increasing number of synoptic surveys is now generating tens to hundreds of transient events per night, and the rates will keep growing, possibly reaching millions of transients per night within a decade or so. Generally, follow-up observations are needed in order to exploit scientifically these data streams to the full. In optical surveys, for instance, all transients look the same when discovered—a starlike object that has changed its brightness significantly—and yet between them they could represent vastly different physical phenomena. Which ones are worthy of a follow-up? This is a critical issue for the massive event streams such as LSST, SKA, etc., and the sheer volume demands an automated approach (Donalek *et al.* 2008; Mahabal *et al.* 2010; Djorgovski *et al.* 2011a).

The process of scientific measurement and discovery operates typically on time-scales from days to decades after the original measurements, feeding back to a new theoretical understanding. However, that clearly will not work when changes occur on time-scales that are shorter than those needed to set up a new round of measurements. It demands real-time systems incorporating a computational analysis and decision engine, and optimized follow-up instruments that can be rapidly deployed with immediate analysis and feedback, and implies automated classification and decision-making systems.

The classification process for a given transient involves: (1) obtaining available contextual archival information, and combining it with the measured parameters from the discovery pipeline, (2) determining (relative?) probabilities or likelihoods of it belonging to some class of transient, (3) obtaining follow-up observations to disambiguate competing classes, (4) using those as a feedback and repeating for an improved classification.

We describe below a few techniques that help in this process. Our principal dataset is the transient event stream from the Catalina Real-time Transient Survey (CRTS; http://crts.caltech.edu; Drake *et al.* 1999; Djorgovski *et al.* 2011b; Mahabal *et al.* 2011), but the methodology we are developing is more universally applicable.

Bayesian Networks. The available data for any given event would generally be heterogeneous and incomplete. That is difficult to accommodate in the standard machine-

learning feature vector approach, but can be naturally accommodated in a Bayesian approach, such as Bayesian Networks (BN) (Mahabal *et al.* 2008).

We have used three colours obtained from the Palomar 60-inch telescope from follow-up observations of CRTS transients, and two contextual parameters: Galactic latitude and proximity to a galaxy. Priors for 6 classes have been used: CVs, SNe, Blazars, other AGN, UV Ceti, and the "Rest" (everything else). We are currently adding more parameters and classes. About 300 objects each have been used for SN and CV, and ∼100 for blazars. The number statistics for other AGN and UV Ceti are still too small. 82% of the objects classified as SN are indeed SN (79% for CVs, 69% for Blazars); the contamination is ∼10–20%. Since a single set of observations achieved that result, the potential for extending the BN and combining its output with other techniques is very promising.

Light-Curve Based Classification. Structure in sparse and/or irregular light curves (LC) can be exploited by automated classification algorithms. Our procedure is to collect LCs for different objects belonging to a given class and representing and encoding the characteristic structure probabilistically in the form of an empirical probability distribution function. That can then be used for subsequent classification of a LC that covers only a few epochs. Moreover, the comparison can be made incrementally over time as new observations become available, the final classification scores improving with each additional set of observations. This forms the basis for a real-time classification methodology. Since the observations come in the form of flux at a given epoch, for each point after the very first one we can form a $(\delta m, \delta t)$ pair. We focus on modelling the joint distribution of all such pairs of data points for a given LC. By virtue of being increments, the empirical probability density functions of these pairs are invariant to absolute magnitude and time shifts—a desirable feature. Upper limits can also be encoded in this methodology, for example, forced photometry magnitudes at a SN location in images taken before the star exploded. We currently use smoothed 2-D histograms to model the distribution of elementary (dm, dt) sets. In our preliminary experimental evaluations with a small number of object classes (single outburst like SNe, periodic variable stars like RR Lyræ and Miras, as well as stochastic variables like blazars and CVs) we have been able to show that the density models for these classes are potentially a powerful method for object classification from sparse/irregular LC data.

Currently we are using the (dm, dt) distributions for classification in a binary mode: successive two-class classifiers in a tree-structure SNe are first separated from non-SNe (the easiest bit, currently performing at ∼99% completness); next, non-SNe are separated into stochastic versus non-stochastic variables, and each group is then further separated into more branches. The most difficult so far has been the CV-blazar node, which is based on just the (dm, dt) density, i.e. without bringing in the proximity to a radio source since we are also interested in discovering blazars that were not active when the archival radio surveys were done. This classifier is currently performing at ∼71% completness. We are also exploring Genetic Algorithms to determine the optimal (dm, dt) bins for different classes. Those will in turn help us optimise follow-up observing intervals for specific classes (Mahabal *et al.* 2011 or Djorgovski *et al.* 2011a).

Incorporating Contextual Information. Contextual information can be highly relevant to resolving competing interpretations; for example, the light curve and observed properties of a transient might be consistent with it being a cataclysmic variable star, a blazar or a SN. If it is subsequently known that there is a galaxy in close proximity, the SN interpretation becomes much more plausible. Such information, however, can be characterized by high uncertainty and absence, and by a rich structure: if there were

https://doi.org/10.1017/S1743921312001056 Published online by Cambridge University Press

two galaxies nearby instead of one then details of galaxy type and structure and native stellar populations become important (for instance, is this type of SN more consistent with it being in the extended halo of a large spiral galaxy or in close proximity to a faint dwarf galaxy?) The ability to incorporate such contextual information in a quantifiable fashion is highly desirable. We have been compiling priors for such information as well; they then get incorporated into the Bayesian network mentioned earlier.

We are also investigating the use of crowd-sourcing (citizen science) as a means of harvesting human pattern-recognition skills, especially in the context of capturing relevant contextual information, and turning them into machine-processable algorithms. A methodology employing contextual knowledge forms a natural extension to the logistic regression and classification methods mentioned above. It will be necessary for larger future surveys where the data flow exceeds available human resources, and moreover it would make such classification objective and repeatable. It also represents an example of a human-machine collaborative discovery process.

Transients can also be found with the technique of image subtraction that employs a matched older observation or a deeper co-added image (Drake *et al.* 1999). If the images are properly matched, a transient stands out as a positive residual, though when used with white light (as is the case with CRTS) the difference images tend to have bipolar residuals, thus leading to false detections. We have been experimenting with this technique to look for supernovæ in galaxies via citizen science, whereby a few amateur astronomers regularly look at the galaxy images along with the residuals presented to them. A large number of SNe have been found in that fashion (see Prieto *et al.* 2011, for an example, and http://nesssi.cacr.caltech.edu/catalina/current.html for a list).

A given classifier may not be optimal for all classes, nor to all types of input, and is the primary reason why multiple types of classifiers have to be employed in the complex task of classifying transients in real time. The presence of different bits of information can trigger different classifiers. In some cases more than one classifier can be used for the same kinds of input. An essential task, then, for handling input from a diverse set of classifiers such as those described above is to derive an optimal event classification. However, combining different classifiers with a different number of output classes and in the presence of error bars is a non-trivial task, and is still under development.

Acknowledgements

This work was supported in part by NASA grant 08-AISR08-0085 and NSF grants AST-0909182 and IIS-1118041.

References

Djorgovski, S. G., *et al.* 2011a, in: A. Srivasatva & N. Chawla (eds.), *Stati. Anal. Data Mining* (CIDU 2011 conf.), in press.

Djorgovski, S. G., *et al.*, 2011b, in: T. Mihara & N. Kawai (eds.), *The First Year of MAXI: Monitoring Variable X-ray Sources* (Tokyo: JAXA Special Publ.), in press

Donalek, C., *et al.* 2008 , in: Bailer-Jones (ed.), *Classification and Discovery in Large Astronomical Surveys*, *AIPC*, 1082, 252,

Drake, A. J., *et al.* 1999, *ApJ*, 521, 602

Drake, A. J., *et al.* 2009, *ApJ*, 696, 870

Mahabal, A. A., *et al.* 2008, *AN*, 329, 3, 288

Mahabal, A. A., *et al.* 2010, *ASPCS*, 434, 115, in: Y. Mizumoto, K. I. Morita & M. Ohishi (eds.), *ADASS XIX*

Mahabal, A. A., *et al.* 2011, *BASI*, 39,387

Prieto, J., *et al.* 2011, *ApJ*, submitted (arXiv:1107.5043)