Future Generation Computer Systems 59 (2016) 95–104

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Real-time data mining of massive data streams from synoptic sky surveys

S.G. Djorgovski [a,*], M.J. Graham [a], C. Donalek [a], A.A. Mahabal [a], A.J. Drake [a], M. Turmon [b], T. Fuchs [b]

[a] California Institute of Technology, Pasadena, CA 91125, USA

[b] Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109, USA

HIGHLIGHTS

- Advances in the automated classification of transient events in synoptic sky surveys.
- Innovative methods for the analysis of irregularly sampled, heterogeneous time series.
- Novel approach to the machine-assisted discovery using a symbolic regression.
- Approaches to an automated decision making based on the automated classification.

ARTICLE INFO

Article history:
Received 16 March 2015
Received in revised form
3 October 2015
Accepted 19 October 2015
Available online 12 January 2016

Keywords:
Sky surveys
Massive data streams
Machine learning
Bayesian methods
Automated decision making

ABSTRACT

The nature of scientific and technological data collection is evolving rapidly: data volumes and rates grow exponentially, with increasing complexity and information content, and there has been a transition from static data sets to data streams that must be analyzed in real time. Interesting or anomalous phenomena must be quickly characterized and followed up with additional measurements via optimal deployment of limited assets. Modern astronomy presents a variety of such phenomena in the form of transient events in digital synoptic sky surveys, including cosmic explosions (supernovae, gamma ray bursts), relativistic phenomena (black hole formation, jets), potentially hazardous asteroids, etc. We have been developing a set of machine learning tools to detect, classify and plan a response to transient events for astronomy applications, using the Catalina Real-time Transient Survey (CRTS) as a scientific and methodological testbed. The ability to respond rapidly to the potentially most interesting events is a key bottleneck that limits the scientific returns from the current and anticipated synoptic sky surveys. Similar challenge arises in other contexts, from environmental monitoring using sensor networks to autonomous spacecraft systems. Given the exponential growth of data rates, and the time-critical response, we need a fully automated and robust approach. We describe the results obtained to date, and the possible future developments.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

The scientific measurement and discovery process traditionally follows the pattern of theory followed by experiment, analysis of results, and then follow-up experiments, often on time scales from days to decades after the original measurements, feeding back to a new theoretical understanding. But that clearly would not work in the case of phenomena where a rapid change occurs on time scales shorter than what it takes to set up the new round of measurements. Thus there is a need for autonomous, real-time scientific measurement systems, consisting of discovery instruments or sensors, a real-time computational analysis and decision engine, and optimized follow-up instruments that can be deployed selectively in (or in near) real-time, where measurements feed back into the analysis immediately. The need for a rapidly analysis, coupled with massive and persistent data streams, implies *a need for an automated classification and decision making.*

This entails some special challenges beyond traditional automated classification approaches, which are usually done in some

* Corresponding author.
E-mail addresses: george@cd3.caltech.edu (S.G. Djorgovski), mjg@cd3.caltech.edu (M.J. Graham), donalek@cd3.caltech.edu (C. Donalek), aam@cd3.caltech.edu (A.A. Mahabal), ajd@cd3.caltech.edu (A.J. Drake), turmon@jpl.nasa.gov (M. Turmon), thomas.fuchs@jpl.nasa.gov (T. Fuchs).

http://dx.doi.org/10.1016/j.future.2015.10.013
0167-739X/© 2016 Elsevier B.V. All rights reserved.

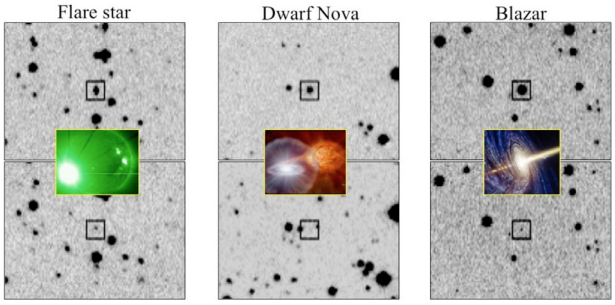

Fig. 1. An illustration of the classification challenge, using examples of transient events from the Catalina Real-time Transient Survey (CRTS) [6–10]. Images in the top row show objects which appear much brighter that night, relative to the baseline images obtained earlier (bottom row). On this basis alone, the three transients are physically indistinguishable, yet the subsequent follow-up shows them to be three vastly different types of phenomena: a flare star (left), a cataclysmic variable powered by an accretion to a compact stellar remnant (middle), and a blazar, flaring due to instabilities in a relativistic jet (right). Accurate transient event classification is the key to their follow-up and physical understanding.

feature vector space, with an abundance of self-contained data derived from homogeneous measurements. The input information here is generally sparse and heterogeneous: there are only a few initial measurements, their types differ from case to case, and the values have differing variances; the contextual information is often essential, and yet difficult to capture and incorporate; many sources of noise, instrumental glitches, etc., can masquerade as transient events; as new data arrive, the classification must be iterated dynamically. There is also the requirement of a high completeness (do not miss any interesting events) and low contamination (not too many false alarms), and the need to complete the classification process and make an optimal decision about expending valuable follow-up resources (e.g., obtain additional measurements using a more powerful instrument, diverting it from other tasks) in real time. These challenges require novel approaches.

Astronomy in particular is facing these challenges in the context of the rapidly growing field of time domain astronomy, based on the new generation of digital synoptic sky surveys that cover large areas of the sky repeatedly, looking for sources that change position (e.g., potentially hazardous asteroids) or change in brightness (a vast variety of variable stars, cosmic explosions, accreting black holes, etc.). Time domain touches upon all subfields of astronomy, from the Solar system to cosmology, and from stellar evolution to the measurements of dark energy and extreme relativistic phenomena. Many important phenomena can be studied only in the time domain (e.g., Supernovae or other types of cosmic explosions), and there is a real possibility of discovering some new, previously unknown types of objects or phenomena.

However, while the surveys discover transient or variable sources, the scientific returns are in their physical interpretation and follow-up observations (Fig. 1). This entails physical classification of objects on the basis of the available data, and an intelligent allocation of limited follow-up resources (e.g., time on other telescopes or space observatories), since generally only a small fraction of all detected events can be followed, and some of them are much more interesting than others. Large data rates and the need for a consistent response imply the need for the automation of these processes, and the problem is rapidly becoming much worse. Today, we deal with data streams of the order of ∼0.1 TB/night and some tens of transients per night; the upcoming Large Synoptic Survey Telescope (LSST) [1] is expected to generate ∼20 TB/night, and millions of transient event alerts. The planned Square Kilometer Array (SKA) [2] radio telescope will move us into the Exascale regime. Thus, a methodology for an automated classification and follow-up prioritization of transient events and variable sources is *critical for the maximum scientific returns* from these planned facilities, in addition to enabling the time domain science now.

In general, most of the major astronomical data sets today are connected and accessible through the Virtual Observatory (VO) framework, which is effectively the global data grid of astronomy [3–5]. However, VO so far does not incorporate many services for knowledge extraction from the massive data sets or data streams, and this is especially important in the context of the time domain astronomy.

The challenges stem from several reasons: first, the data are sparse, especially right after the initial detection; archival and contextual information is essential (e.g., the spatial context of the source, the multi-wavelength context, and the temporal context — have the source been detected before, and if so, what was its variability behavior, etc.). Both the subsequent measurements (if any) and the archival information are likely to be highly heterogeneous and/or incomplete. The probabilistic classification of the events evolves as new data arrive, and is used to generate priorities and automated follow-up decisions and requests, which are then feed back into the system. However, the availability of the follow-up resources also changes in time, which affects their value, and is limited in allocation; etc.

To respond to these challenges, we have been developing and testing a variety of automated classification approaches for time domain astronomy. We divided the problem into two parts: event classification, and follow-up recommendations given the available assets. Our preliminary results have been described, e.g., in [11–21]. Here we give some updates to these papers and some of our current work. For additional reviews and references, see, e.g., [22–28].

As a testbed development data stream, we use transient events and variable sources discovered by the Catalina Real-Time Transient Survey (CRTS) [6–10]. CRTS provides a great variety of physical object types, and a realistic heterogeneity and sparsity of data. We found that a number of published methods, developed on "de luxe" data sets, to say nothing about the simulated data, simply fail or significantly underperform when applied to the more realistic data (in terms of the cadences, S/N, seasonal modulation, etc.), typified by the CRTS data stream. In general, we find that every method has some dependence on the quantity and quality of the input data (e.g., the number of measurements in a light curve, the sampling strategy, etc.), and all of our tests incorporate assessment of the robustness and applicability of a given method in different data regimes.

Whereas our focus is on an astronomical context, similar situations arise in many other fields, where anomalies or events of interest must be identified in some massive data stream, characterized, and responded to in as close to the real time as possible (e.g., environmental monitoring, security, etc.).

2. Bayesian networks

Bayesian techniques may be the most promising approach for the classification with sparse, incomplete, or missing data, since, generally speaking, one can use the information from the available priors, regardless of what data are not available. In particular, we experimented with a Bayesian Network (BN) [29] based classifier, as it offers a natural way of incorporating a variety of the measurements of different types, and more can be added as they become available. However, the network complexity increases super-exponentially as more variables are included, and there is a premium of selecting a small number of the most powerful classification discriminating features (see below).

Our initial implementation used follow-up measurements of photometric colors obtained at the Palomar 60-inch telescope. For example, in the relative classification of Cataclysmic Variables

Download English Version:

https://daneshyari.com/en/article/424861

Download Persian Version:

https://daneshyari.com/article/424861

Daneshyari.com