Automated Real-Time Classification and Decision Making in Massive Data Streams from Synoptic Sky Surveys

S. G. Djorgovski*, A. A. Mahabal, C. Donalek,
M. J. Graham, A. J. Drake

California Institute of Technology
Pasadena, CA 91125, USA
[george,aam,donalek,mjg,ajd]@astro.caltech.edu

M. Turmon, T. Fuchs

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA 91109, USA
[turmon, thomas.fuchs]@jpl.nasa.gov

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 arise 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.

Keywords-classification; sky surveys; massive data streams; machine learning; Bayesian methods; automated decision making

I. 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 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 (don't 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. 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.

To respond to these challenges, we have been developing and testing a variety of automated classification approaches for time domain astronomy. Our preliminary results have been described, e.g., in [3,4,5,6,7,8,9,10,11,12]. Here we give some updates to these papers and some of our current work. For additional reviews and references, see, e.g., [13,14,15, 16,17,18,19].

As a testbed development data stream, we use transient events and variable sources discovered by the Catalina Real-Time Transient Survey (CRTS) [20,21,22,23,24]. 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 may 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.).

II. 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) [25] 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.

Our initial implementation used measurements of photometric colors obtained at the Palomar 60-inch telescope. For example, in the relative classification of Cataclysmic Variables (CVs) vs. Supernovae, we obtain a completeness of ~ 80% and a contamination of ~ 19%.

We found BN to be an excellent way of incorporating quantitative spatial contextual information, e.g., the proximity of a given transient event to the nearest star or the nearest galaxy detected in the Sloan Digital Sky Survey (SDSS) [26]. For example, a transient (nearly) coincident with a galaxy will most likely be a Supernova (SN), whereas a transient coincident to with a star-like object in an archival survey such as the SDSS would more likely be some type of a variable star or an Active Galactic Nucleus (AGN). These are limited by the depth and the angular resolution of the comparison archival survey, but for our tests, transients from CRTS and comparisons with SDSS are well matched for this purpose.
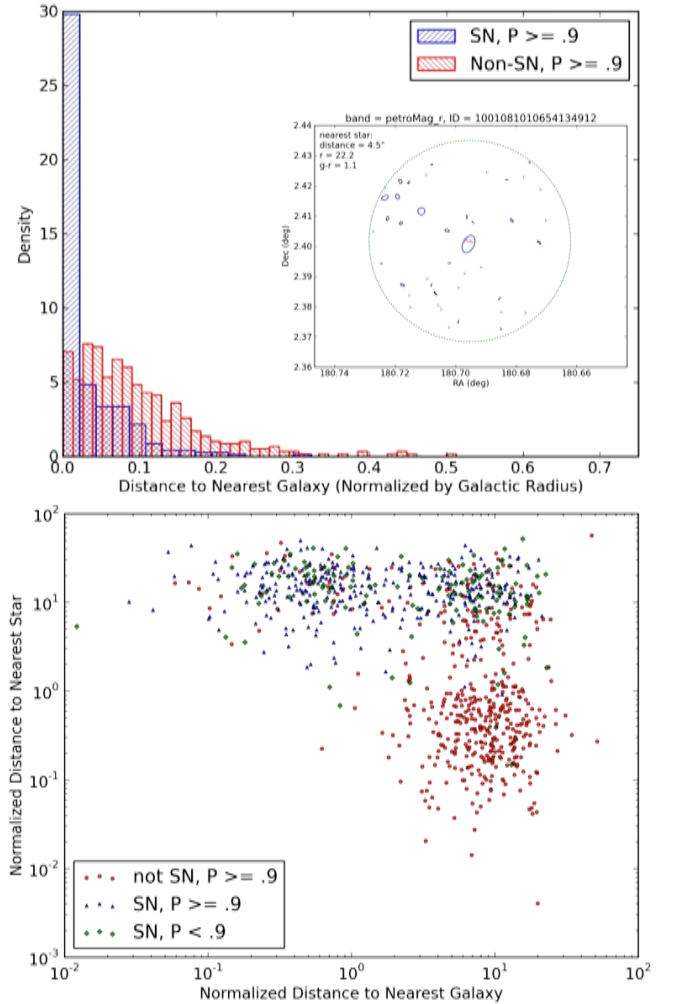

Figure 1. Top: The distribution of normalized distances to the nearest galaxy (using the Petrosian radius metric, see the text) for the transients classified as having a probability of > 90% of being SNe (blue), and having a probability of > 90% of *not* being SNe (red). The inset shows a distribution of objects in the field of a particular transient, with galaxies represented as ellipses, scaled by their magnitude. Bottom: The distribution of these kinds of objects, plus those classified as possible SNe, but with a probability < 90%, in the parameter space defined by the two distances.

In the case of proximity to the nearest star, a simple angular distance is sufficient. In the case of galaxies, an ambiguity arises: is a closer, but very faint galaxy more likely to be the possible SN host, or a considerably brighter galaxy that is a little further away? Thus, a different metric is needed, and we use angular separation in the units of characteristic radii for the light distribution in galaxies. After some experimentation, we decide on the so-called Petrosian radius, which is one of the parameters provided by the SDSS archive.

Temporal contextual information is also important. Another distinguishing characteristic of SNe is that they can explode only once, so a presence of previously detected spikes in the light curve of a given transient diminishes the likelihood of it being a SN.

Thus, we construct a BN with 3 input variables, the proximity to the nearest star, to the nearest galaxy (suitably normalized), and a light curve peak statistic developed by us. The results are illustrated in Fig. 1. The results using just these 3 contextual variables (the nearest star distance, the nearest galaxy distance, and the peak statistics) are very encouraging. For the transients correctly classified as SNe, the completeness is in the range ~ 80% – 92% with contamination in the range ~ 18% – 29%. For the transients correctly classified as *not* being SNe, the completeness is in the range ~ 79% – 83% with contamination in the range ~ 8% – 14%. These results can be improved substantially by introducing other priors, e.g., colors, or light curve based parameters, at the expense of an increased computational complexity.

III. STATISTICAL DESCRIPTORS OF VARIABILITY AND THE OPTIMAL FEATURE SELECTION

Data heterogeneity is perhaps the key problem for the automated classification of astronomical light curves, or, for that matter, any other irregularly sampled time series. Since the numbers of the data points and their temporal separations vary, the light curves themselves cannot be used directly in any method that assumes data in the form of uniform feature vectors. In order to circumvent this problem, we evaluate a number of statistical descriptors of light curves that can be evaluated regardless of the number of data points or the cadence, e.g., the variance of the observed magnitudes, the skew, kurtosis, etc. About 60 such parameters have been defined in the literature, to which we added a dozen of our own devising. Their definitions can be found at the *Caltech Time Series Characterization Service* [27]. These statistical descriptors can then be used to form feature vectors that can be fed into automated classifiers.

Obviously, not all would be equally useful, and different ones may be more useful in different circumstances. We are conducting a detailed study of their utility for different aspects of the classification problem, for different classifiers, and in different data regimes (e.g., S/N, number of data points, etc.). Ideally, one seeks combinations of features that optimally separate different classes of transients or variables. Here we summarize some of the key results; more details are given in [28].
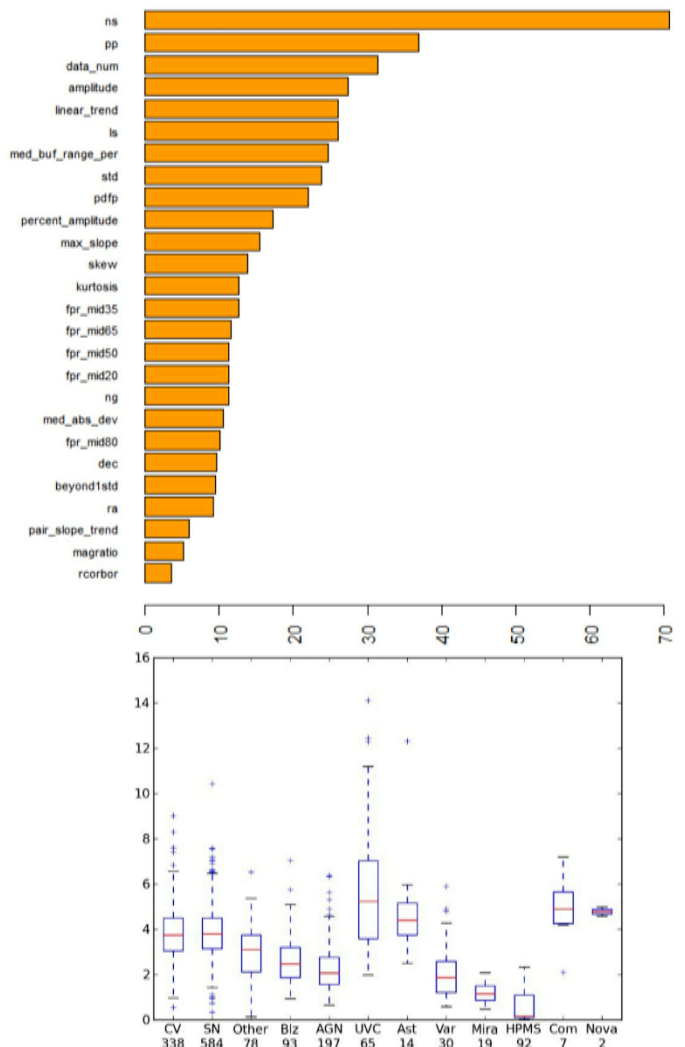

Figure 2. Top: A relative ranking of light curve features in terms of the classification discriminating power. Bottom: Standard box plot representation of the distributions of the top ranked parameter (*nsigma*) for different physical types of transients and variables.

Given a set of feature vectors, a broad variety of automated classification tools can be applied, both supervised and unsupervised. Supervised methods include artificial neural networks (ANN), and in particular the multi-layer perceptron (MLP), support vector machines (SVM), decision trees (DT) and their generalization random forests (RF), etc. Unsupervised methods include Kohonen self-organizing maps (SOM), k Means (KM), k nearest neighbors (kNN), etc. Given a particular classifier, and a particular classification problem, e.g., separating two different types of periodic variables, or supernovae vs. non-explosive transients, we can evaluate the relative importance of different features using several methods.

One way to reduce the dimensionality of the input space is applying a forward feature selection strategy that consists in selecting a subset of features from the training set that best predict the test data by sequentially selecting features until there is no improvement in prediction [35,36]. The optimal feature selection varies both with the particular classification

problem (e.g., separating two different types of variable stars) and the algorithm used. We have performed an extensive set of experiments for this optimization.
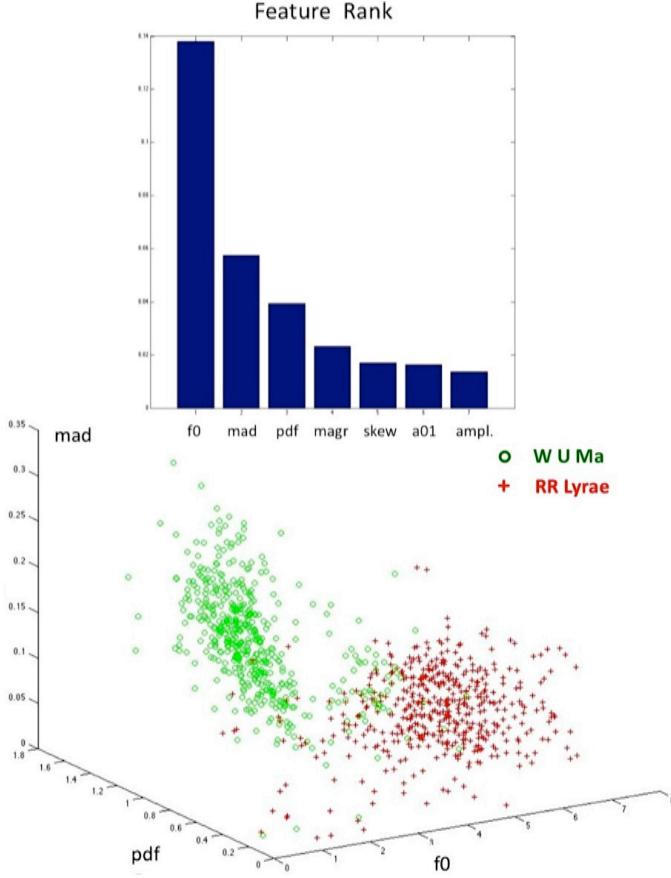

Figure 3. Top: A relative ranking of light curve features for a particular classification problem in separating two types of variable stars, RR Lyrae (red crosses) and W UMa (green circles). Bottom: The two classes are separated very effectively in the 3-dimensional space of the 3 top ranked features.

We have employed different classifiers in the selected feature space to assess the performance of different feature selection algorithms, to prove that feature selection strategies actually help in reduce the dimensionality of the problem without loss in accuracy. The performance of the classifiers is rated based on the following three criteria. *Completeness* is the percentage of objects of a given class correctly classified as such. *Contamination* is the percentage of objects of a given class, incorrectly classified as belonging to another class. *Loss* is the fraction of misclassified data.

To avoid overfitting, a cross-validation approach is recommended, e.g., with 10-fold cross-validation the original sample is randomly partitioned into 10 subsamples. Each time a single subsample is retained as test data, and the remaining are used as training data. This process is then repeated 10 times with each of the subsamples used exactly once as the test. In presence of few training data, Leave-One-Out Cross-Validation (LOOCV) may be used: a single observation from the original sample is used as the validation data, and the remaining observations as the training data. This is repeated

such that each observation in the sample is used once as the validation data. Leave-one-out cross-validation is usually very computationally expensive because of the large number of times the training process is repeated.

For example in an experiment where we classify two types of variable stars, RR Lyrae and W UMa, using a Relief method, only four parameters out of the 60 available were selected (Fig. 3). For this particular problem, we obtain completeness rates of ~ 96-97%, and contamination rates of ~ 3-4%. It is interesting to note that the parameters automatically selected by this procedure essentially represent the period-amplitude relationship which is used to differentiate between subclasses of RR Lyrae, i.e., the algorithm uncovers a physically meaningful relation.

A more challenging, but more realistic and relevant problem is multi-class classification. To find the parameters that give the most of the classification discriminating information, we have used a subset from CRTS containing six classes (Supernovae, Cataclysmic Variables, Blazars, other AGNs, RR Lyrae and Flare Stars) and 20 parameters. Table 1 shows some of these results for two different multi-class experiments. It is interesting to note that different features appear among the most significant subsets, depending on the physical nature of the classes considered.

Class	Complet	Contam	Features
Blazar	81%	13%	Amplitude, linear_trend, flux_percentile_ratio_mid20, percent_diff_flux_percentile, qso, skew, std, stetson_j , stetson_k, lomb-scargle
CV	96%	5%	
RR Lyrae	97%	5%	
SN Ia	99%	1%	

Class	Complet	Contam	Features
Blazar	83%	13%	Amplitude, beyond1std, flux_percentile_ratio_mid65, max_slope, qso, std, lomb-scargle
CV	94%	6%	
RR Lyrae	97%	4%	

Table 1. Optimal feature selection in two different multi-class experiments.

Thus, we see that feature selection strategies can lead to a substantial dimensionality reduction and improved classifier performance in a broad range of astrophysical situations.

Since many types of variable stars show a periodic behavior, and periods and their significance play an important role in their physical classification, we conducted a detailed study of different algorithms for period determination [29]. We find that superior results are obtained using the Conditional Entropy algorithm [30]. More details are given in these references.

IV. MACHINE-ASSISTED DISCOVERY

As the exponential growth of data volumes, rates, and complexity continues, we may see an increased use of methods for a collaborative human-computer discovery. Recognizing meaningful patterns and correlations in high dimensionality data parameter spaces is a very non-trivial task.

Another novel approach that we explored in the course of this study is the use of Machine Discovery, i.e., software that can formulate and test data models. The particular package that we used, with M. Graham as the lead, is *Eureqa* [31]. Here we outline some of the key results; more details are given in [32].

Eureqa is a software tool which aims to describe a data set by identifying the simplest mathematical formulae which could describe the underlying mechanism that produced the data. It employs symbolic regression to search the space of mathematical expressions to determine the best-fitting functional form – this involves fitting both the form of the equation and its parameters simultaneously. Binary classification can be cast as a problem amenable to this tool – the "trick" is to formulate the search relationship as: *class = g(f(x_1, x_2, x_3, ..., x_n))* where *g* is either the Heaviside step function or the logistic function, which gives a better search gradient. Eureqa finds a best-fit function, *f*, to the data that will get mapped to a 0 or a 1, depending on whether it is positively or negatively valued (or lies on either side of a specified threshold, say 0.5, in the case of the logistic function.)

We considered three specific binary light curve classification problems using *Eureqa*: RR Lyrae vs. W UMa (Fig. 4), CV vs. blazar, and Type Ia vs. core-collapse Supernovae. For each case, we compiled data sets of light curves from the CRTS survey for the appropriate classes of objects, and derived ~30 – 60 dimensional feature vectors for each object. A set of 10 *Eureqa* runs was performed for each case with each run omitting 10% of the data and the best-fit solution for that run then applied with the omitted data as the validation set so giving us 10x-cross-validation on the resulting solutions.

For example, in the binary classification of these periodic variables, *Eureqa* correctly identifies the optimal feature parameter plane that separates them as physically distinct classes (Figs. 5, 6). This is very impressive, since the program does not "know" anything about these objects, and simply discovers the relationship contained in the data.

Some of the preliminary results for multiple classes, comparing *Eureqa* with one of the best "traditional" machine learning methods, Decision Trees (DT), are given in Table 2. We note that DT is a supervised classification method, and thus it incorporates the domain knowledge from the training data set; *Eureqa* has no such expert-provided input. Even so, the results are broadly comparable for most classes. *Eureqa* does not do as well in the situation where the light curves are qualitatively similar, e.g., blazars vs. CVs, or different subtypes of Supernovae. However, a random person with no expert knowledge in this field (just as *Eureqa* doesn't have it) would probably also fail completely in separating those classes.
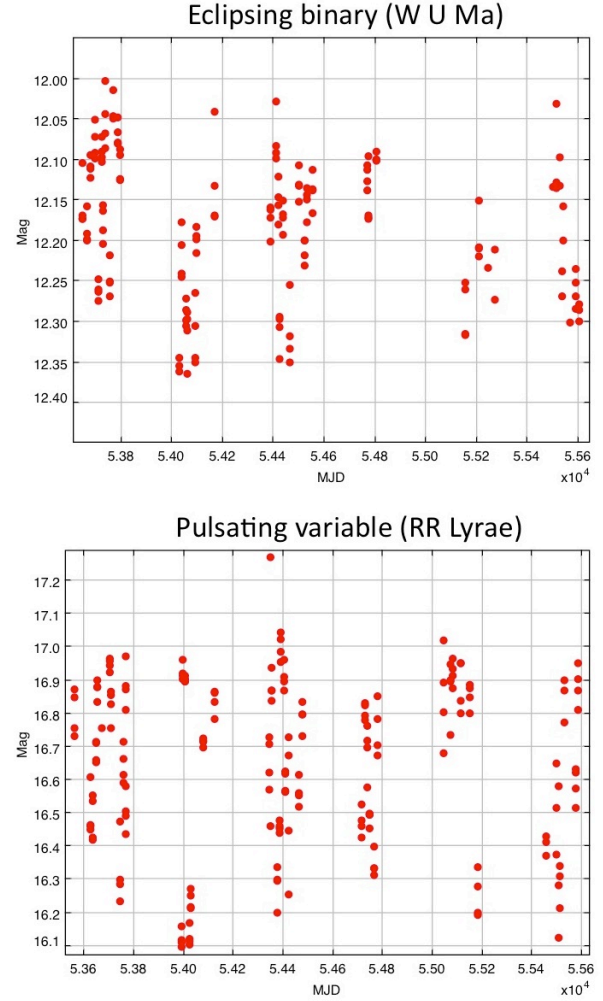

Figure 4. Light curves of two types of periodic variables (not folded by the period) from the CRTS survey, used in this experiment.
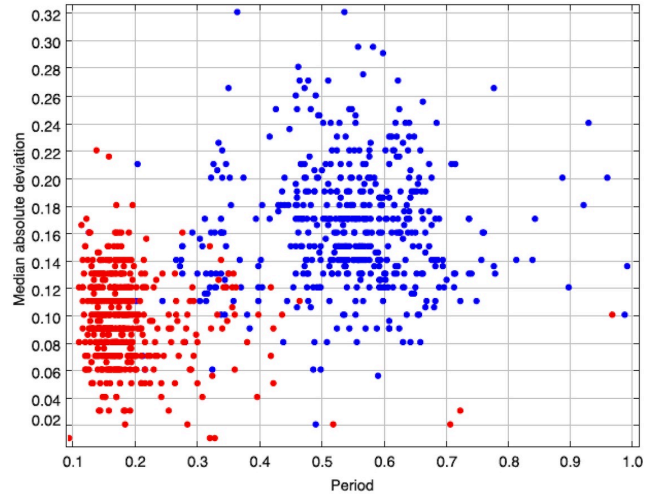

Figure 5. Separation of two types of periodic variables, RR Lyrae (blue) and W UMa (red) in the optimal feature plane discovered by *Eureqa*.

Data set	EUREQA		Decision tree	
	Purity (per cent)	Efficiency (per cent)	Purity (per cent)	Efficiency (per cent)
RR Lyrae	98	96	95	95
W UMa	97	99	96	96
CV	89	91	92	92
Blazar	68	63	87	83
SN Ia	76	93	90	96
CC SN	74	41	92	80

Table 2. Performance of Eureqa compared with that of a traditional DT classifier for several classes of variable objects. From [REF].

As these preliminary results show, at least in some cases *Eureqa* can identify and characterize physically meaningful structures in feature vector data to a sufficient degree that it can be employed for binary classification. An advantage of this is that *Eureqa* provides an analytical expression to separate the classes rather than relying on application of a trained black box algorithm. We see this as one of the first steps in a practical *human-computer collaborative discovery* in the era of big data. We think that such novel methods will become increasingly important for the data-intensive science in the 21st century.

V. METACLASSIFICATION: OPTIMAL COMBINING OF CLASSIFIERS AND CONTEXTUAL KNOWLEDGE

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 both it being a cataclysmic variable star, an active galactic nucleus, or a supernova. If it is subsequently known that there is a galaxy in close proximity, the supernova interpretation becomes much more plausible. Such information, however, can be characterized by high uncertainty and absence, and by a rich structure – if there were two candidate host galaxies, their morphologies, distance, etc., become important, e.g., is this type of supernova more consistent with 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 are investigating the use of crowdsourcing as a means of harvesting human pattern recognition skills, especially in the context of capturing the relevant contextual information, and turning it into machine-processable algorithms.

We can identify three possible sources of information that can be used to find the unknown parameters. They can be from *a priori* knowledge, e.g. from physics or monotonicity considerations, or from examples that are labeled by experts, or from the feedback from downstream observatories once labels are determined. The first case would serve to give an analytical form for the distribution, but the second two amount to the provision of labeled examples, (x, y), which can be used to select a set of k probability distributions.

A methodology employing contextual knowledge forms a natural extension to the logistic regression and classification methods mentioned above. Ideally such knowledge can be expressed in a manipulable fashion within a sound logical model, for example, it should be possible to state the rule that "a supernova has a stellar progenitor and will be substantially brighter than it by several order of magnitude" with some metric of certainty and infer the probabilities of observed data matching it.

Markov Logic Networks (MLN) [33] are such a probabilistic framework using declarative statements (in the form of logical formulae) as atoms associated with real-valued weights expressing their strength. The higher the weight, the greater the difference in log probability between a world that satisfies the formula and one that does not, all other things being equal. In this way, it becomes possible to specify 'soft' rules that are likely to hold in the domain, but subject to exceptions - contextual relationships that are likely to hold such as supernovae may be associated with a nearby galaxy or objects closer to the Galactic plane may be stars.

The structure of a MLN – the set of formulae with their respective weights – is also not static but can be revised or extended with new formulae either learned from data or provided by third parties. In this way, new information can easily be incorporated. Continuous quantities, which form much of astronomical measurements, can also be easily handled with a hybrid MLN. This approach could be used to represent a set of different classifiers and the inferred most probable state of the world from the MLN would then give the optimal classification.
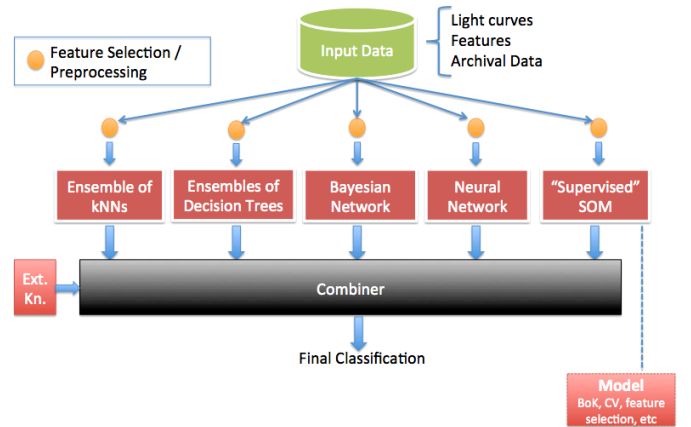

Figure 12. A schematic illustration of the metaclassifier for an optimal combination of the output of different classifiers.

We are also experimenting with the "sleeping expert" method [34]. A set of different classifiers each generally works best with certain kinds of inputs. Activating these optionally only when those inputs are present provides an optimal solution to the fusion of these classifiers. Sleeping expert can be seen as a generalization of the *if-then* rule: *if* this condition is satisfied *then* activate this expert, e.g., a

specialist that makes a prediction only when the instance to be predicted falls within their area of expertise. For example, some classifiers work better when certain inputs are present, and some work only when certain inputs are present. It has been shown that this is a powerful way to decompose a complex classification problem. External or *a priori* knowledge can be used to awake or put experts to sleep and to modify online the weights associated to a given classifier; this contextual information may be also expressed in text.

VI. Classification-Informed Automated Decision Making

While at least preliminary astrophysical classifications of variable sources and transient events may be obtained using survey and archival data and the methods described above, in many cases the classifications will be ambiguous, or, in the case of particularly interesting events, additional data from other instruments would be needed to fully exploit them scientifically. This poses the challenge of automated decision making as to the optimal use of the available, finite follow-up resources, e.g., other telescopes or instruments. This work is still in progress, but we outline here some of the key ideas.

We typically have sparse observations of a given object of interest, leading to classification ambiguities among several possible object types (e.g., when an event is roughly equally likely to belong to two or more possible object classes, or when the initial data are simply inadequate to generate a meaningful classification at all). Generally speaking, some of them would be of a greater scientific interest than others, and thus their follow-up observations would have a higher scientific return. Observational resources are scarce, and always have some cost function associated with them, so a key challenge is to determine the follow-up observations that are most useful for improving classification accuracy, and detect objects of scientific interest.

There are two parts to this challenge. First, what type of a follow-up measurement – given the *available* set of resources (e.g., only some telescopes/instruments may be available) – would yield the maximum information gain in a particular situation? And second, if the resources are finite and have a cost function associated with them (e.g., you can use only so many hours of the telescope time), when is the potential for an interesting discovery worth spending the resources?

We take an information-theoretic approach to this problem that uses Shannon entropy to measure ambiguity in the current classification. We can compute the entropy drop offered by the available follow-up measurements – for example, the system may decide that obtaining an optical light curve with a particular temporal cadence would discriminate between a supernova and a flaring blazar, or that a particular color measurement would discriminate between, say, a cataclysmic variable eruption and a gravitational microlensing event. A suitable prioritized request for the best follow-up observations would be sent to the appropriate robotic (or even human-operated) telescopes.

Alternatively, instead of maximizing the classification accuracy, we consider a scenario where the algorithm chooses a set of events for follow-up and subsequent display to an astronomer. The astronomer then provides information on how interesting the observation is. The goal of the algorithm is to learn to choose follow-up observations which are considered most interesting. This problem can be naturally modeled using *Multi-Armed Bandit* algorithms (MAB). The MAB problem can abstractly be described as a slot machine with k levers, each of which has different expected returns (unknown to the decision maker). The aim is to determine the best strategy to maximize returns. There are two extreme approaches: (1) exploitation – keep pulling the lever which, as per your current knowledge, returns most, and (2) exploration – experiment with different levers in order to gather information about the expected returns associated with each lever. They key challenge is to trade off exploration and exploitation. There are algorithms guaranteed to determine the best choice as the number of available tries goes to infinity.

In this analogy different telescopes and instruments are the levers that can be pulled. Their ability to discriminate between object classes forms the returns. This works best when the priors are well assembled and a lot is already known about the type of object one is dealing with. But due to the heterogeneity of objects, and increasing depth leading to transients being detected at fainter levels, and more examples of relatively rarer subclasses coming to light, treating the follow-up telescopes as a MAB will provide a useful way to rapidly improve the classification and gather more diverse priors. An analogy could be that of a genetic algorithm which does not get stuck in a local maxima because of its ability to sample a larger part of the parameter space.

VII. Concluding Comments

Our goal in this paper was to illustrate the richness and the challenges associated with the problem of an automated classification of transient events and variable sources (or, more generally, heterogeneous time series of measurements of a population of objects containing a number of different classes). Whereas this is one of the core challenges of the vibrant and emerging field of time-domain astronomy, similar problems can be easily identified in other domains.

Several aspects of this problem make it particularly interesting: dealing with the data heterogeneity and sparsity; use of statistical descriptors to form feature vectors, instead of using the data directly; dimensionality reduction of feature spaces that is context-dependent; forays into the collaborative human-computer discovery; optimal combining of different classifiers that is also context dependent; and finally, optimal allocation of limited follow-up resources when there are multiple cost functions involved.

REFERENCES

[1] LSST: http://www.lsst.org/

[2] SKA: http://www.skatelescope.org/

[3] A. Mahabal et al. (PQ survey team), "Automated probabilistic classification of transients and variables", Astonomische Nachrichten, 329, 288, 2008.

[4] A. Mahabal, P. Wozniak, C. Donalek and S.G. Djorgovski, "Transients and Variable Stars in the Era of Synoptic Imaging", in: LSST Science Book, eds. Z. Ivezic, et al., Ch. 8.4, p. 261; available at http://www.lsst.org/lsst/scibook, 2009.

[5] A. Mahabal et al., "Mixing Bayesian Techniques for Effective Real-time Classification of Astronomical Transients", in: Proc. ADASS XIX, ed. Y. Mizumoto, ASP Conf. Ser., 434, 115, 2010.

[6] A. Mahabal et al., "Classification of Optical Transients: Experiences from PQ and CRTS Surveys", in: Gaia: At the Frontiers of Astrometry, eds. C. Turon, et al., EAS Publ. Ser. 45, 173, Paris: EDP Sciences, 2010.

[7] A. Mahabal et al., "The Meaning of Events", in: Hotwiring the Transient Universe, eds. S. Emery Bunn, et al., Lulu Enterprises Publ. http://www.lulu.com/, p. 31, 2010.

[8] A. Mahabal et al., "Towards Real-Time Classification of Astronomical Transients", AIP Conf. Ser., 1082, 287, 2008.

[9] C. Donalek et al., "New Approaches to Object Classification in Synoptic Sky Surveys", AIP Conf. Ser., 1082, 252, 2008.

[10] S. G. Djorgovski, et al., "Towards an Automated Classification of Transient Events in Synoptic Sky Surveys", Proc. CIDU 2011, eds. A. Srivasatva, et al., NASA publ., p. 174, 2011.

[11] S. G. Djorgovski, et al., "Flashes in a Star Stream: Automated Classification of Astronomical Transient Events", Proc. e-Science 2012, IEEE press, 2012.

[12] A.A. Mahabal, et al., "Real Time Classification of Transient Events in Synoptic Sky Surveys", Proc. IAU Symp. 285, New Horizons in Time Domain Astronomy, eds. E. Griffin et al., p. 355. Cambridge: Cambridge Univ. Press (2012)

[13] S.G. Djorgovski, A.A. Mahabal, A.J. Drake, M.J. Graham, C. Donalek, C., "Sky Surveys", in: Astronomical Techniques, Software, and Data (ed. H. Bond), Vol. 2 of Planets, Stars, and Stellar Systems (ser. ed. T. Oswalt), p. 223. Berlin: Springer Verlag, 2013.

[14] S.G. Djorgovski, A.A. Mahabal, A.J. Drake, M.J. Graham, C. Donalek, C., R. Williams, "Exploring the Time Domain With Synoptic Sky Surveys", Proc. IAU Symp. 285, New Horizons in Time Domain Astronomy, eds. E. Griffin et al., p. 141. Cambridge: Cambridge Univ. Press, 2012.

[15] M.J. Graham, et al., "Connecting the Time Domain Community with the Virtual Astronomical Observatory", in: Observatory Operations: Strategies, Processes and System IV, eds., Peck, A., Seaman, R., Comeron, F., Proc. SPIE, 8448, 84480P, 2013.

[16] M.J. Graham, et al., "Data Challenges of Time Domain Astronomy", in Distrib. Parallel Databases, eds. Qiu, X., Gannon, D., Vol. 30 (5-6), 371-384, 2012.

[17] J. Bloom, et al., "Automating Discovery and Classification of Transients and Variable Stars in the Synoptic Survey Era", to appear in Publ. Astron. Soc. Pacific, 2011.

[18] J. Richards, et al., "On Machine-learned Classification of Variable Stars with Sparse and Noisy Time-series Data", Astrophys. J., 733, 10, 2011.

[19] J. Bloom and J. Richards, "Data Mining and Machine-Learning in Time-Domain Discovery & Classification", in: Advances in Machine Learning and Data Mining for Astronomy, in press; arXiv/1104.3142, 2011.

[20] Catalina Sky Survey (CRTS): http://crts.caltech.edu/

[21] A.J. Drake, et al., "First Results from the Catalina Real-time Transient Survey", Astrophys. J., 696, 870, 2009.

[22] S.G. Djorgovski, A. Drake et al. (the CRTS survey team), "The Catalina Real-Time Transient Survey (CRTS)", in The First Year of MAXI: Monitoring Variable X-ray Sources, eds. T. Mihara & N. Kawai, Tokyo: JAXA Special Publ., 2011.

[23] A.A. Mahabal, et al., "Discovery, classification, and scientific exploration of transient events from the Catalina Real-Time Transient Survey", Bull. Astron. Soc. India, 39, 38, 2011.

[24] A.J. Drake, et al., "The Catalina Real-time Transient Survey", Proc. IAU Symp. 285, New Horizons in Time Domain Astronomy, eds. E. Griffin et al., p. 306. Cambridge: Cambridge Univ. Press , 2012.

[25] D. Heckerman, "A Tutorial on Learning with Bayesian Networks", in: Learning in Graphical Models, ed. M. Jordan, Cambridge, MA: MIT Press, 1999.

[26] Sloan Digital Sky Survey (SDSS): http://sdss.org

[27] *Caltech Time Series Characterization Service*, http://nirgun.caltech.edu:8000/

[28] C. Donalek, et al., "Feature Selection Strategies for Classifying High Dimensional Astronomical Data Sets", in: Scalable Machine Learning: Theory and Applications, IEEE BigData 2013.

[29] M.J. Graham, et al., "A comparison of period finding algorithms", Monthly Notices Royal Astron. Soc., 434, 3423–3444, 2013.

[30] M.J. Graham, et al., "Using conditional entropy to identify periodicity", Monthly Notices Royal Astron. Soc., 434, 2629–2635, 2013.

[31] Schmidt M., Lipson H., "Distilling free-form natural laws from experimental data", Science, 324, 81-85, 2009.

[32] M.J. Graham, S.G. Djorgovski, A.A. Mahabal, C. Donalek, A.J. Drake, "Machine-assisted discovery of relationships in astronomy", Monthly Notices Royal Astron. Soc., 431, 2371–2384, 2013.

[33] M. Richardson and P. Domingos, "Markov logic networks", Machine Learning, 62, 107-136, 2006.

[34] Y. Freund, R. Schapire, Y. Singer, M. Warmuth, "Using and combining predictors that specialize", in: STOC '97 Proceedings of the twenty-ninth annual ACM symposium on theory of computing", 334-343, New York: ACM, 1997.

[35] R.S. Marko, K. Igor, "Theoretical and empirical analysis of Relief and ReliefF," Machine Learning Journal, 53:23-69, 2003.

[36] Y. Sun, J. Li, "Iterative Relief for feature weighting", ICML '06 Proc. 23rd International Conference on Machine Learning, 913-920. New York: ACM, 2006.