A Machine Learning Classification Broker for Petascale Mining of Large-scale Astronomy Sky Survey Databases

Kirk D. Borne
Department of Computational & Data Sciences, George Mason University
kborne@gmu.edu

Abstract

We describe the new data-intensive research paradigm that astronomy and astrophysics is now entering. This is described within the context of the largest data-producing astronomy project in the coming decade – the LSST (Large Synoptic Survey Telescope). The enormous data output, database contents, knowledge discovery, and community science expected from this project will impose massive data challenges on the astronomical research community. One of these challenge areas is the rapid machine learning, data mining, and classification of all novel astronomical events from each 3-gigapixel (6-GB) image obtained every 20 seconds throughout every night for the project duration of 10 years. We describe these challenges and a particular implementation of a classification broker for this data fire hose.

1. Introduction

The development of models to describe and understand scientific phenomena has historically proceeded at a pace driven by new data. The more we know, the more we are driven to tweak or to revolutionize our models, thereby advancing our scientific understanding. This data-driven modeling and discovery linkage has entered a new paradigm [1]. The acquisition of scientific data in all disciplines is now accelerating and causing a nearly insurmountable data avalanche [2]. In astronomy in particular, rapid advances in three technology areas (telescopes, detectors, and computation) have continued unabated – all of these advances lead to more and more data [3]. With this accelerated advance in data generation capabilities, humans will require novel, increasingly automated, and increasingly more effective scientific knowledge discovery systems [4].

To meet the data-intensive research challenge, the astronomical research community has embarked on a grand information technology program, to describe and unify all astronomical data resources worldwide. This global interoperable virtual data system is referred to as the National Virtual Observatory (NVO) in the U.S., or more simply the “Virtual Observatory” (VO). Within the international research community, the VO effort is steered by the International Virtual Observatory Alliance (IVOA). This grand vision encompasses more than a collection of data sets. The result is a significant evolution in the way that astrophysical research, both observational and theoretical, is conducted in the new millennium [5]. This revolution is leading to an entirely new branch of astrophysics research – *Astroinformatics* – still in its infancy, consequently requiring further research and development as a discipline in order to aid in the data-intensive astronomical science that is emerging [6].

The VO effort enables discovery, access, and integration of data, tools, and information resources across all observatories, archives, data centers, and individual projects worldwide [7]. However, it remains outside the scope of the VO projects to generate new knowledge, new models, and new scientific understanding from the huge data volumes flowing from the largest sky survey projects [8, 9]. Even further beyond the scope of the VO is the ensuing feedback and impact of the potentially exponential growth in new scientific knowledge discoveries back onto those telescope instrument operations. In addition, while the VO projects are productive science-enabling I.T. research and development projects, they are not specifically scientific research projects. There is still enormous room for scientific data portals and data-intensive science research tools that integrate, mine, and discover new knowledge from the vast distributed data repositories that are now VO-accessible [4].

The problem therefore is this: astronomy researchers will soon (if not already) lose the ability to keep up with any of these things: the data flood, the scientific discoveries buried within, the development of new models of those phenomena, and the resulting new data-driven follow-up observing strategies that are imposed on telescope facilities to collect new data needed to validate and augment new discoveries.

2. Astronomy Surveys as Data Producers

A common feature of modern astronomical sky surveys is that they are producing massive (terabyte) databases. New surveys may produce hundreds of terabytes (TB) up to 100 (or more) petabytes (PB) both in the image data archive and in the object catalogs (databases). Interpreting these petabyte catalogs (i.e., mining the databases for new scientific knowledge) will require more sophisticated algorithms and networks that discover, integrate, and learn from distributed petascale databases more effectively.

2.1. The LSST Sky Survey Database

One of the most impressive astronomical sky surveys being planned for the next decade is the Large Synoptic Survey Telescope project (LSST, http://www.lsst.org/) [10]. The three fundamental distinguishing astronomical attributes of the LSST project are:

(1) *Repeated temporal measurements* of all observable objects in the sky, corresponding to thousands of observations per each object over a 10-year period, expected to generate 10,000-100,000 alerts each night – an alert is a signal (e.g., XML-formatted RSS feed) to the astronomical research community that something has changed at that location on the sky: either the brightness or position of an object, or the serendipitous appearance of some totally new object;

(2) W*ide-angle imaging* that will repeatedly cover most of the night sky within 3 to 4 nights (= tens of billions of objects); and

(3) *Deep co-added images* of each observable patch of sky (summed over 10 years: 2014-2024), reaching far fainter objects and to greater distance over more area of sky than other sky surveys [11].

Compared to other astronomical sky surveys, the LSST survey will deliver time domain coverage for orders of magnitude greater number of objects. It is envisioned that this project will produce ~30 TB of data per each night of observation for 10 years. The final image archive will be ~60 PB, and the final LSST astronomical object catalog (object-attribute database) is expected to be ~10-20 PB.

2.2. The LSST Data-Intensive Science Challenge

LSST is not alone. It is one (likely the biggest one) of several large astronomical sky survey projects beginning operations now or within the coming decade. LSST is by far the largest undertaking, in terms of duration, camera size, depth of sky coverage, volume of data to be produced, and real-time requirements on operations, data processing, event-modeling, and follow-up research response. One of the key features of these surveys is that the main telescope facility will be dedicated to the primary survey program, with no specific plans for follow-up observations. This is emphatically true for the LSST project [12]. Paradoxically, the follow-up observations are scientifically essential – they contribute significantly to new scientific discovery, to the classification and characterization of new astronomical objects and sky events, and to rapid response to short-lived transient sky phenomena.

Since it is anticipated that LSST will generate many thousands (probably tens of thousands) of new astronomical event alerts per night of observation, there is a critical need for innovative follow-up procedures. These procedures necessarily must include modeling of the events – to determine their classification, time-criticality, astronomical relevance, rarity, and the scientifically most productive set of follow-up measurements. Rapid time-critical follow-up observations, with a wide range of time scales from seconds to days, are essential for proper identification, classification, characterization, analysis, interpretation, and understanding of nearly every astrophysical phenomenon (e.g., supernovae, novae, accreting black holes, microquasars, gamma-ray bursts, gravitational microlensing events, extrasolar planetary transits across distant stars, new comets, incoming asteroids, trans-Neptunian objects, dwarf planets, optical transients, variable stars of all classes, and anything that goes “bump in the night”).

2.3. Petascale Data Mining with the LSST

LSST and similar large sky surveys have enormous potential to enable countless astronomical discoveries. Such discoveries will span the full spectrum of statistics: from rare one-in-a-billion (or one-in-a-trillion) type objects, to a complete statistical and astrophysical specification of a class of objects (based upon millions of instances of the class). One of the key scientific requirements of these projects therefore is to learn rapidly from what they see. This means: (a) to identify the serendipitous as well as the known; (b) to identify outliers (e.g., “front-page news” discoveries) that fall outside the bounds of model expectations; (c) to identify rare events that our models say should be there; (d) to find new attributes of known classes; (e) to provide statistically robust tests of existing models; and (f) to generate the vital inputs for new models. All of this requires integrating and mining of all known data: to train classification models and to apply classification models.

LSST alone is likely to throw such data mining and knowledge discovery efforts into the petascale realm. For example: astronomers currently discover ~100 new supernovae (exploding stars) per year. Since the beginning of human history, perhaps ~10,000 supernovae have been recorded. The identification, classification, and analysis of supernovae are among the key science requirements for the LSST Project to explore Dark Energy – i.e., supernovae contribute to the analysis and characterization of the ubiquitous cosmic Dark Energy. Since supernovae are the result of a rapid catastrophic explosion of a massive star, it is imperative for astronomers to respond quickly to each new event with rapid follow-up observations in many measurement modes (light curves; spectroscopy; images of the host galaxy's environment). Historically, with <10 new supernovae being discovered each week, such follow-up has been feasible. But now, LSST promises to produce a list of 1000 new supernovae each night for 10 years [11], which represent a small fraction of the total (10-100 thousand) alerts expected each night! Astronomers are faced with the enormous challenge of efficiently mining, correctly classifying, and intelligently prioritizing a staggering number of new events for follow-up observation each night for a decade.

3. A Classification Broker for Astronomy

We are beginning to assemble user requirements and design specifications for a machine learning engine (data integration network plus data mining algorithms) to address the petascale data mining needs of the LSST and other large data-intensive astronomy sky survey projects. The data requirements surpass those of the current Sloan Digital Sky Survey by 1000-10,000 times, while the time-criticality requirement (for event/object classification and characterization) drastically drops from months down to minutes (or tens of seconds). In addition to the follow-up classification problem (described above), astronomers also want to find every possible new scientific discovery (pattern, correlation, relationship, outlier, new class, etc.) buried within these new enormous databases. This might lead to a petascale data mining compute engine that runs in parallel alongside the data archive, testing every possible model, association, and rule. What we are focusing on here is the time-critical data mining engine (i.e., classification broker) that enables rapid follow-up science for the most important and exciting astronomical discoveries of the coming decade, on a wide range of time scales from seconds to days, corresponding to a plethora of exotic astrophysical phenomena.

3.1. Broker Specifications: AstroDAS

The classification broker's primary specification is to produce and distribute scientifically robust near-real-time classification of astronomical sources, events, objects, or event hosts. These classifications are derived from integrating and mining data, information, and knowledge from multiple distributed data repositories. The broker feeds off existing robotic telescope and astronomical alert networks world-wide, and then integrates existing astronomical knowledge (catalog data) from the VO. The broker may eventually provide the knowledge discovery and classification service for LSST, a torrential fire hose of data and astronomical events.

Incoming event alert data will be subjected to a suite of machine learning (ML) algorithms for event classification, outlier detection, object characterization, and novelty discovery. Probabilistic ML models will produce rank-ordered lists of the most significant and/or most unusual events. These ML models (e.g., Bayesian networks, decision trees, multiple weak classifiers, Markov models, or perhaps scientifically derived similarity metrics) will be integrated with astronomical taxonomies and ontologies that will enable rapid information extraction, knowledge discovery, and scientific decision support for real-time astronomical research facility operations – to follow up on the 10-100K alertable astronomical events that will be identified each night for 10 years by the LSST sky survey.

The classification broker will include a knowledgebase to capture the new labels (tags) that are generated for the new astronomical events. These tags are annotations to the events. "Annotation" refers to tagging the data and metadata content with descriptive terms. For this knowledgebase, we envision a collaborative tagging system, called AstroDAS (Astronomy Distributed Annotation System) [13]. AstroDAS is similar to existing science knowledgebases, such as BioDAS [14], WikiProteins [15], the Heliophysics Knowledgebase (HPKB) [16], and The Entity Describer [17]. AstroDAS is "distributed" in the sense that the source data and metadata are distributed, and the users are distributed. "Annotation" refers to tagging the data and metadata content with descriptive terms, which apply to individual data granules or to subsets of the data. It is a "system" with a unified schema for the annotation database, where distributed data are perceived as a unified data system to the user. One possible implementation of AstroDAS could be as a Web 2.0 (=Science2.0) mashup. AstroDAS users will include providers (authors) and annotation users (consumers). Consumers (humans or machines) will eventually interact with AstroDAS in four ways:

1. Integrate the annotation database content with their own data portals.
2. Subscribe to receive notifications when new sources are annotated or classified.
3. Use the classification broker as a data integration tool to broker classes and annotations between sky surveys, robotic telescopes, and data repositories.
4. Query the annotation database (either manually or through web services).

In the last case, the users include the astronomical event message producers, who will want to issue their alerts with their best-estimate for the astronomical classification of their event. The classification will be generated through the application of machine learning algorithms to the networked data accessible via the VO, in order to arrive at a prioritized list of classes, ordered by probability of certainty.

3.2. Collaborative Annotation of Classes

Machine learning and data mining algorithms, when applied to very large data streams, can generate the classification labels (tags) autonomously. Generally, scientists do not want to leave this decision-making to machine intelligence alone – they prefer to have human intelligence in the loop also. When humans and machines work together to produce the best possible classification label(s), this is collaborative annotation. Collaborative annotation is a form of Human Computation [18]. Human Computation refers to the application of human intelligence to solve complex difficult problems that cannot be solved by computers alone. Humans can see patterns and semantics (context, content, and relationships) more quickly, accurately, and meaningfully than machines. Human Computation therefore applies to the problem of annotating, labeling, and classifying voluminous data streams. Of course, the application of autonomous machine intelligence (data mining and machine learning) to the annotation, labeling, and classification of data granules is also valid and efficacious. The combination of both human and machine intelligence is critical to the success of AstroDAS as a classification broker for enormous data-intensive astronomy sky survey projects, such as LSST. Figure 1 highlights the main components of AstroDAS.

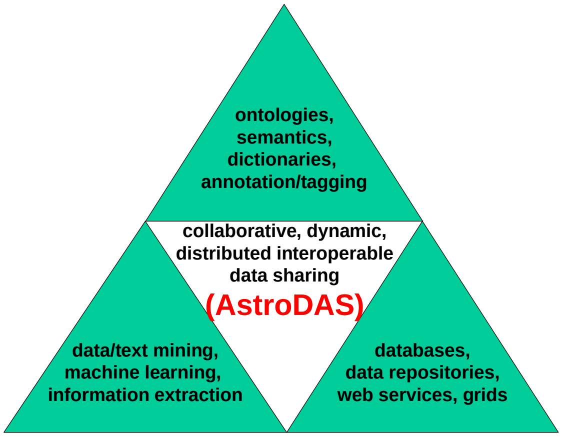
ontologies, semantics, dictionaries, annotation/tagging
collaborative, dynamic, distributed interoperable data sharing
(AstroDAS)
data/text mining, machine learning, information extraction
databases, data repositories, web services, grids

Figure 1. Main components of AstroDAS.

4. A Research Agenda

We identify some of the key research activities that must be addressed, in order to promote the development of a machine learning-based classification broker for petascale mining of large-scale astronomy sky survey databases. Many of these research activities are already being pursued by other data mining and computational science researchers – we hope to take advantage of all such developments, many of which are enabled through advanced next-generation data mining and cyber-infrastructure research:

- Before the classification labels can be useful, we must reach community consensus on the correct set of semantic ontological, taxonomical, and classification terms. There are ontologies under development in astronomy already – their completeness, utility, and usability need to be researched.
- Research into user requirements and scientific use cases will be required in order that we design, develop, and deploy the correct user-oriented petascale data mining system.
- A complete set of classification rules must be researched and derived for all possible astronomical events and objects. For objects and events that are currently unknown, we need to identify robust outlier and novelty detection rules and classifiers. These need to be researched and tested.
- We need to research and collect comprehensive sets of training examples for the numerous classes that we hope to classify. With these samples, the classification broker will be trained and validated.
- Algorithms for web services-based (perhaps grid-based or peer-to-peer) classification and mining of distributed data must be researched, developed, and validated. These mining algorithms should include text mining as well as numeric data mining, perhaps an integrated text-numeric data mining approach will be most effective and thus needs to be researched.
- User interface and interaction models will need to be researched through prototypes and demonstrations of the classification broker.
- Research into the robust integration of the many system components identified in Figure 1 will be needed. This will require investigation of different modes of interaction and integration, such as grids, web services, RSS feeds, ontologies (expressed in RDF or OWL), linked databases, etc.
- Deploy a working classification broker on a live astronomical event message stream, to research its functionality, usefulness, bottlenecks, failure modes, security, robustness, etc. Fortunately, there are such event message feeds available today, though on a much smaller scale than that anticipated from LSST.

Clearly, this is an ambitious research agenda. It will not be fully accomplished in just a year or two. It will require several years of research and development. This is fortunate, since the most dramatic need for the classification broker system for astronomy will come with the start-up of LSST sky survey operations in 2014, lasting ten years (until 2024). So, we have a few years to get it right, and we will need all of those years to complete the challenging research program described above.

5. Summary: Astroinformatics

Finally, we close with discussions of BioDAS (the inspiration behind AstroDAS) and of the relevance of informatics (e.g., Bioinformatics and Astroinformatics) to the classification broker described in this paper. Informatics is the discipline of organizing, accessing, mining, analyzing, and visualizing data for scientific discovery. Another definition says "informatics is the set of methods and applications for integration of large datasets across spatial and temporal scales to support decision-making, involving computer modeling of natural systems, heterogeneous data structures, and data-model integration as a framework for decision-making" [19].

Massive scientific data collections impose enormous challenges to scientists: how to find the most relevant data, how to reuse those data, how to the mine data and discover new knowledge in large databases, and how to represent the newly discovered knowledge. The bioinformatics research community is already solving these problems with BioDAS (Biology Distributed Annotation System) [14]. The DAS provides a distributed system for researchers anywhere to annotate (mark-up) their own knowledge (tagged information) about specific gene sequences. Any other researcher anywhere can find this annotation information quickly for any gene sequence. Similarly, astronomers can annotate individual astronomical objects with their own discoveries. These annotations can be applied to observational data/metadata within distributed digital data collections. The annotations provide mined knowledge, class labels, provenance, and semantic (scientifically meaningful) information about the experiment, the experimenter, the object being studied (astronomical object in our case, or gene sequence in the case of the bioinformatics research community), the properties of that object, new features or functions discovered about that object, its classification, its connectiveness to other objects, and so on.

Bioinformatics (for biologists) and Astroinformatics (for astronomers) provide frameworks for the curation, discovery, access, interoperability, integration, mining, classification, and understanding of digital repositories through (human plus machine) semantic annotation of data, information, and knowledge. We are focusing on further development of Astroinformatics as: (1) a new subdiscipline of astronomical research (similar to the role of bioinformatics and geoinformatics as stand-alone subdisciplines in biological and geoscience research and education, respectively); and (2) the new paradigm for data-intensive astronomy research and education, which focuses on existing cyberinfrastructure (such as the National Virtual Observatory). This integrated research and education activity matches well to the objectives of the new NSF CDI (Cyber-enabled Discovery and Innovation) initiative [20] and the new CODATA ADMIRE (Advanced Data Methods and Information technologies for Research and Education) initiative [21].

6. References

[1] Mahootian, F., & Eastman, T. 2007, "Complementary Frameworks of Scientific Inquiry: Hypothetico-Deductive, Hypothetico-Inductive, and Observational-Inductive," *World Futures* journal, in press.

[2] Bell, G., Gray, J., & Szalay, A. 2005, "Petascale computations systems: Balanced cyberinfrastructure in a data-centric world," downloaded from http://arxiv.org/abs/cs/0701165.

[3] Becla, J., et al. 2006, "Designing a multi-petabyte database for LSST," downloaded from http://arxiv.org/abs/cs/0604112.

[4] Borne, K. D. 2006, "Data-Driven Discovery through e-Science Technologies," in the proceedings of the IEEE conference on Space Mission Challenges for I.T.

[5] McDowell, J. C. 2004, "Downloading the Sky", IEEE Spectrum, 41, p. 35.

[6] Borne, K. D., & Eastman, T. 2006, "Collaborative Knowledge-Sharing for E-Science," in the proceedings of the AAAI conference on Semantic Web for Collaborative Knowledge Acquisition.

[7] Plante, R., et al. 2004, "VO Resource Registry," in the proceedings of the ADASS XIII conference, downloaded on August 23, 2007 from http://www.us-vo.org/pubs/index.cfm.

[8] Borne, K. D. 2001a, "Science User Scenarios for a VO Design Reference Mission: Science Requirements for Data Mining," in "Virtual Observatories of the Future," p.333.

[9] Borne, K. D. 2001b, "Data Mining in Astronomical Databases," in "Mining the Sky," p.671.

[10] Tyson, J. A. 2004, "The Large Synoptic Survey Telescope: Science & Design," downloaded on August 23, 2007 from http://www.lsst.org/Meetings/CommAccess/abstracts.shtml.

[11] Strauss, M. 2004, "Towards a Design Reference Mission for the LSST," downloaded on August 23, 2007 from http://www.lsst.org/Meetings/CommAccess/abstracts.shtml.

[12] Mould, J. 2004, "LSST Followup," downloaded from http://www.lsst.org/Meetings/CommAccess/abstracts.shtml on August 23, 2007.

[13] Bose, R., Mann, R., & Prina-Ricotti, D. 2006. "AstroDAS: Sharing Assertions across Astronomy Catalogues through Distributed Annotation," Lecture Notes in Computer Science (LNCS) Volume 4145, 193-202.

[14] http://biodas.org/

[15] http://www.wikiprofessional.info/

[16] http://www.lmsal.com/helio-informatics/hpkb/

[17] Good, B., Kawas, E., & Wilkinson, M. 2007, "Bridging the gap between social tagging and semantic annotation: E.D. the Entity Describer," downloaded on September 26, 2007 from http://precedings.nature.com/documents/945/version/2 .

[18] von Ahn, L., & Dabbish, L. 2004, "Labeling Images with a Computer Game,“ in the proceedings of the SIGCHI conference on Human Factors in Computing Systems, p.319.

[19] Downloaded on August 23, 2007 from http:// ag.arizona.edu/srnr/research/wr/breshears/informatics_UA .

[20] http://www.nsf.gov/news/news_summ.jsp?cntn_id=108366

[21]www.iucr.org/iucr-top/data/docs/codataga2006_beijing.html