

SkyDOT (Sky Database for Objects in the Time Domain): A Virtual Observatory for Variability Studies at LANL

P. Wozniak, K. Borozdin, M. Galassi, W. Priedhorsky, D. Starr, W. T. Vestrand, R. White
and
J. Wren

Los Alamos National Laboratory, Los Alamos, NM, USA

ABSTRACT

The mining of Virtual Observatories (VOs) is becoming a powerful new method for discovery in astronomy. Here we report on the development of SkyDOT (Sky Database for Objects in the Time domain), a new Virtual Observatory, which is dedicated to the study of sky variability. The site will confederate a number of massive variability surveys and enable exploration of the time domain in astronomy. We discuss the architecture of the database and the functionality of the user interface. An important aspect of SkyDOT is that it is continuously updated in near real time so that users can access new observations in a timely manner. The site will also utilize high level machine learning tools that will allow sophisticated mining of the archive. Another key feature is the real time data stream provided by RAPTOR (RAPid Telescopes for Optical Response), a new sky monitoring experiment under construction at Los Alamos National Laboratory (LANL).

Keywords: Virtual Observatory, variable stars, database, real-time sky monitoring, Data Mining

1. INTRODUCTION

The past decade astronomy has seen the advent of numerous data intensive projects. In the early nineties, microlensing experiments, which daily monitored tens of millions of objects over long periods of time, were pushing the limits of commonly available computer storage and processing power (e.g. Ref. 1). Typically microlensing teams implemented their own specialized database systems with very limited portability. As a result, development effort was often duplicated. Nevertheless, the scientific payoff of those projects went far beyond the primary goal, that is the discovery and study of microlensing events. The wealth of data created an information rich environment, where serendipitous science happens continuously in studies of stellar populations in the dense fields toward the Galactic Bulge and Magellanic Clouds.²

Microlensing searches provided an unprecedented record of variability in those selected fields, with hundreds of epochs for each object gathered over the several year baseline. However, no such record has been published for most of the sky. In fact, the bright sky between 6 and 15 mag is largely unexplored in terms of variability.³ There is even less data for astronomical events at short time scales, that have to be identified in real time in order to be studied. Automated online data analysis with alert capability is required for success in this domain. The list of active projects measuring positions and brightness over significant parts of the sky includes more than 30 names (<http://www.astro.princeton.edu/faculty/bp.html>). Only a few of these manage to process the data timely and make it available to all astronomers in a useful form. Data overload seems to be a frequent occurrence in astronomy today.

The concept of the Virtual Observatory⁴ promises to solve many of the problems with very large data sets. It is recognized that increasing the amount of available data by an order of magnitude, and considering joint multi-wavelength, spatial and temporal information from multiple surveys simultaneously, opens up a new discovery space. Variability studies are essential to numerous astronomical questions, however the time dimension adds even more information to be processed. With the use of modern database systems and emerging standards for

Send correspondence to P. Wozniak:

E-mail: wozniak@lanl.gov, Telephone: 1 505 667 1381,

Address: Los Alamos National Laboratory, MS-D436, Los Alamos, NM 87545, USA

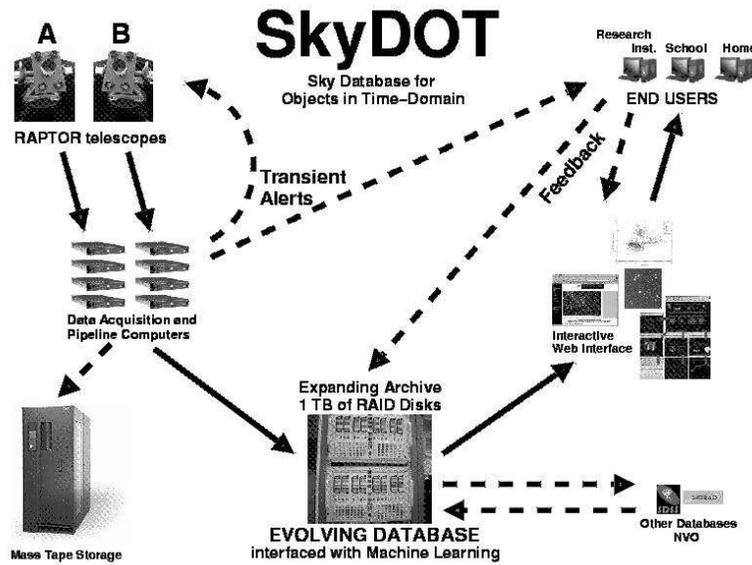


Figure 1. The concept of SkyDOT, A Virtual Observatory for Variability Studies at LANL. A real time data stream from RAPTOR will be integrated with the database. The main data flow is clockwise from the upper left to the upper right corner, that is from data acquisition hardware all the way to the broad user. Tape backup and interfaces with other databases are also shown. Important parts of the design are closed loop operation of RAPTOR telescopes, real time transient alerts and feedback collecting.

web interfaces, numerous data sets can be federated, despite the fact that differing technologies may be used at different nodes.^{5,6} Data Mining and Machine Learning are becoming very attractive tools for extracting knowledge from vast quantities of data.

The Sloan Digital Sky Survey (SDSS) is an example of a project where technical issues of the efficient data distribution were given serious consideration,⁷ although the SDSS SkyServer Database does not provide the temporal data. There are several project that do have time domain data and are working on making those available to the astronomical community.⁸ To the best of our knowledge, none of the teams (including the authors of this contribution) can provide the full sky coverage with prompt online data access. This paper describes the work in progress at the Los Alamos National Laboratory to build a Virtual Observatory for studies of variable objects across the sky.

2. INTERACTIVE ASTRONOMICAL VARIABILITY DATABASE AT LANL

2.1. Data Sources

The concept of the Virtual Observatory for studies of variable objects is illustrated in Fig. 1. Los Alamos National Laboratory (LANL) is involved in several sky monitoring projects. It is planned that the sky variability database we are constructing will consist of multiple data sets, all converted to similar formats and available through a common user interface with easy to apply graphical tools. Some of the current data sets that will be included are public OGLE-II data, RAPTOR and ROTSE. One of the authors developed a Difference Image Analysis pipeline for the Optical Gravitational Lensing Experiment and processed OGLE-II Galactic Bulge data (Ref. 9 and references therein). RAPid Telescopes for Optical Response (RAPTOR) is a new, stereoscopic search for optical transients in real time.¹⁰ The Robotic Optical Transient Search Experiment (ROTSE) is a GRB followup project responding to satellite triggers.¹¹ Mirrors of other publicly available synoptic data sets will also be incorporated over time. Below we summarize briefly what to expect from each of the data sets.

2.1.1. OGLE

OGLE-II survey¹² was conducted with the 1.3 meter Warsaw telescope at the Las Campanas Observatory, Chile. The portion of the data collected during observing seasons 1997–1999 has been analyzed using the Difference Image Photometry, approximately calibrated to a standard system, and is available in public domain.⁹ Only variable objects have been measured with this technique. Between 200 and 300 *I*-band frames are available for each of the 49 OGLE-II Galactic Bulge fields. The number of detected variable objects per field varies between 800 and 9000 due to variations of the stellar density and uneven frequency of observations. Each field covers 14×57 arcmin, for a total of ~ 11 square deg. The range of covered galactic longitudes is roughly ± 11 deg. The database comprises a total of over 51×10^6 individual photometric records for 221,801 objects with *I*-band magnitudes between 10.5 and 20.0. The rate of spurious objects is still about 10%.

2.1.2. RAPTOR

RAPTOR^{10, 13} is a new generation optical transient search at LANL, NM. Its key features are real time data analysis, closed loop operation with rapid slewing and response to interesting events, and stereo vision for high confidence rejection of artifacts. Each of the two identical RAPTOR telescopes, separated by 38 km, will have four 85 mm cameras with the 1500 square deg. total field of view, and a central, more sensitive 400 mm camera with much narrower, 2×2 deg. field of view. An 0.3 m Ritchey-Chretien telescope with a transmission grating will provide low resolution spectroscopy for selected objects. The data rate from all imaging instruments will reach 4 TB/year. Fast cadence time histories will be constructed for roughly 300,000 objects across all locally visible sky up to the limiting magnitude 12.5. With the additional sky patrol instrument, RAPTOR experiment can also cover all locally visible sky to about 16 mag (about 30 million objects) in about 2–3 nights. The main challenge for the RAPTOR database will be real time operations. All data will be available in public domain as soon as technically possible, preferably as a real time update to the online database.

2.1.3. ROTSE

For almost 4 years the ROTSE-I telescope nightly patrolled all the sky visible from Los Alamos, NM. Those observations constitute a valuable database for studying the variability of the sky in the 8–15.5 magnitude range. For the purpose of an all-sky variability census¹⁴ we are using the data taken between April 1999 and March 2000. Observations were performed in 640 fields, each covering 8×8 deg. The most difficult part of data processing, reducing images to object lists, is complete. This data set amounts to more than 225,000 wide field images totaling approximately 2.5 TB of data. Photometry alone takes 250 GB of binary storage. The number of available observations varies between about 300 and 40 near declinations $+90$ deg and -30 deg respectively. Roughly 32,000 periodic variables are expected to be found based on the scaling of a pilot study. The total number of objects with time histories is over 2×10^7 , and the total number of individual photometric measurements reaches 3.5×10^9 .

2.2. Logical Database Design

Logical design is concerned with the general data model and data structures for the database, independently of any hardware issues or the selected Database Management System (DBMS). An Object Oriented model is natural wherever there is a need for grouping parameters and their transformations together (classes). However, performance problems were reported for a particular Object Oriented Database Management System (OODBMS) (Ref. 7 and references therein). A relational system proved more successful in that case. We adopt a relational data model (see Ref. 15 for a very accessible presentation of theory). The problem of constructing a database of temporal variability has certain similarities to a so called data warehouse. It resembles a full record of all customer receipts in a supermarket rather more than inventory database for the same store. Object parameters and events are recorded and accumulated in succession, usually without many further updates. This is the major difference with respect to a transaction oriented system, where concurrent reads and updates of previous values are essential. Instead of fluctuating around a fixed size, the database content tends to grow at a significant rate. Even if we consider only a fixed amount of data, the structure of chained events remains. Fig. 2 shows the schema we developed for the ROTSE-I database. With minor modifications this design will handle many other data sets. The backbone is a main event table containing all available measurements

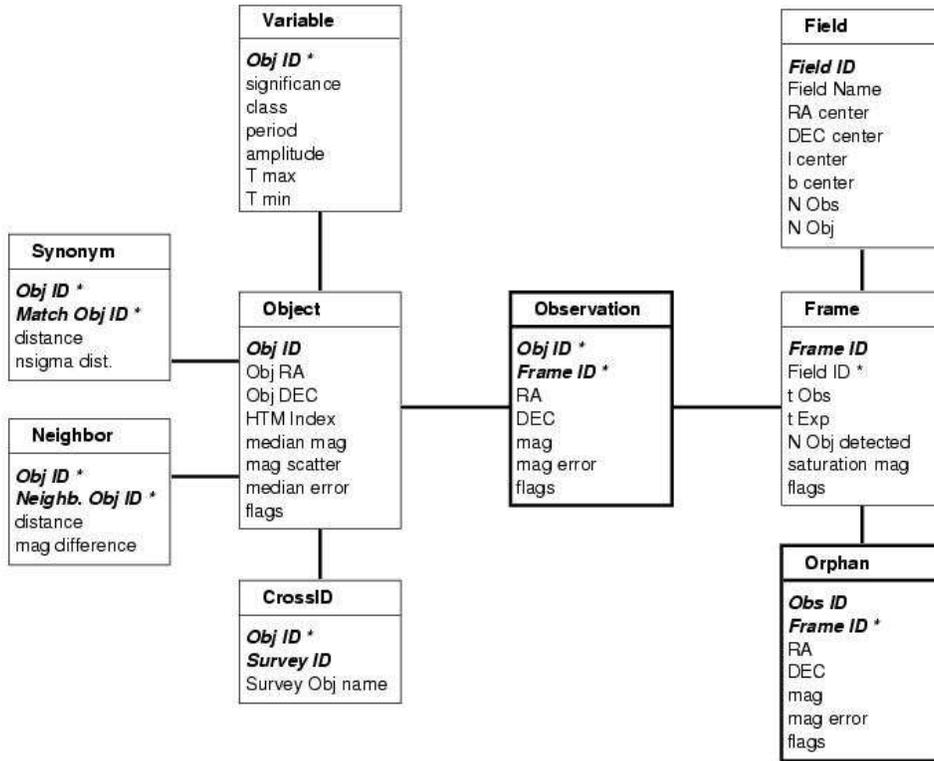


Figure 2. The relational schema of the variability database under construction. Event tables are shown with thick lines. Bold slanted font indicates primary key attributes and asterisk indicates a foreign key. Various parts of the snowflake surrounding the main fact table serve as entry points for most queries.

(**Observation**). Each measurement is identified by a composite primary key that includes the frame number and the object number. These identifiers are foreign keys which connect two smaller tables called dimension tables: **Object**, **Frame**. Dimension tables are basically entry points into the database. Most queries start from those surrounding tables and are then read from a portion of the main photometry table. To maximize the speed at the expense of data redundancy and update anomalies, the dimension tables are often highly denormalized, resulting in a star schema. We decided against this approach and expanded the design into a snowflake schema with no unnecessary data redundancy.

We expect cone searches (a “circle” in RA and DEC) to be the most popular use of the database. For that purpose the **Object** table contains a Hierarchical Triangle Mesh (HTM) index allowing fast extraction of objects in requested areas of the sky (Section 2.4). Full photometry can be obtained from **Observation** table for thus selected objects. A separate table contains information specific to variable objects and the number of entries in **Variable** table is only a few percent of the number of all objects. Variables are also flagged in the **Object** table (we model “is type of” relationship with a foreign key in a subclass table). The fact that any given object can be detected in more than one field is a significant complication. In such cases there may be multiple object IDs associated with a single physical object. We take this into account by providing a list of synonymous pairs of object IDs. The advantage of this approach is that cross identification of multiple references to the same object can be revised with very little effort as more is known about the data. In a similar way we precompute the information on nearest neighbors, providing a valuable diagnostic tool for blending related problems. The database objects should be cross identified with other surveys. **CrossID** table in Fig. 2 is just a starting point for thorough cross referencing of different surveys.

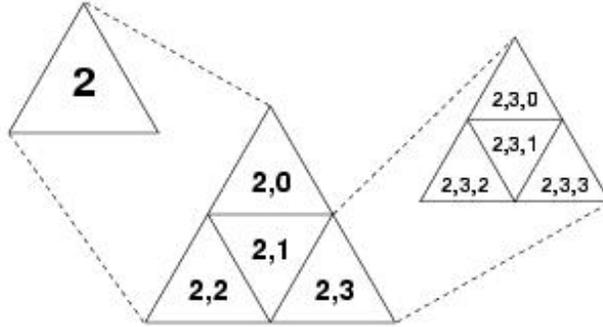


Figure 3. The idea of the Hierarchical Triangle Mesh (after Kunszt et al.). A recursive procedure assigns a number to any point on a sphere. Spatial queries use this index to limit searches to a relatively small number of triangles with minimum possible area.

On the other side of the main fact table, the **Frame** table provides the data on individual measurements and allows selection of photometry based on exposure parameters and observing conditions. Julian Dates of observations are also stored there and do not have to be duplicated for all objects. Position dependent effects like heliocentric time correction are not always required and can be calculated on the fly. The **Frame** is basically in “part of” relationship with **Field** because of the observing protocols in synoptic surveys. A primary key of **Field** is posted in **Frame** to reflect this fact. Preprocessing of the data and construction of positional templates on field by field basis implies association of objects with certain fields (complications occur in the overlap regions). Ideally, such associations should be removed and all objects should be treated independently of the field they came from. In practice, full merging of the data is difficult before all systematic effects are removed from photometry and possibly astrometry as well. This is another reason to keep the **Field** table in the database: diagnostic purposes and tracking systematics.

The detections unidentified with any of the template objects are normally present in every exposure. We store such measurements in **Orphan** table for further analysis at a later time. Those detections can be spurious, but in many cases they are real and of great interest, possibly being associated with moving objects or transients.

2.3. Physical Considerations

As the basis of our database system we employ PostgreSQL DBMS (<http://www.postgresql.org>), the most advanced open source system, which is available for free and has a very relaxed license. PostgreSQL has numerous object oriented features added on top of a relational system. Formally it handles unlimited amount of data with some minor restrictions like the 64 TB maximum table size before the need to split. PostgreSQL has transaction support for real time applications like tracking/updating states of interesting objects and alert systems. For a fixed size data set of ROTSE-I we can take maximal approach to indexing, that is define indexes on all primary and foreign keys and most of the non-floating-point arguments. It remains to be seen whether frequent rebuilding a large number of indexes in a system with real-time updates is still an acceptable overhead. We are developing a web browser based GUI that will be a primary means of connecting to the database, both internally at LANL and externally over the Internet. The user interface and most of the database application layer will be based on ADOdb (<http://php.weblogs.com/ADODB>), a database connectivity library written in PHP. ADOdb is very portable as it supports numerous DBMSs on several platforms and all SQL calls are independent of the particular database system, provided an appropriate driver is loaded.

The main data storage is currently on two 1 GHz Pentium III RedHat Linux boxes, each with 512 MB of 133 MHz RAM and 600 GB of raw EIDE disk space. Under RAID 5 this gives 1.1 TB of total available space. Both machines are connected to a 100 baseT ethernet line. Because of the concerns over security and continuous availability, eventually the entire database will be fully replicated and only one of the copies will be accessible from outside the local intranet.

2.4. Spatial Querying

Until recently RDBMS typically did not include built in structures for fast querying of spatial data. Positional searches require such functionality for clustering studies, correlation functions and quick cross identification of objects. The solution is to construct an external 2D or 3D index, store it as a column in the object table, and further index it using one of the standard methods within the database. Following the SDSS Sky Server experience,⁶ we adopted the Hierarchical Triangular Mesh (HTM) code^{16,17} developed at Johns Hopkins University. The HTM code employs a set of canonical transformations that project the sphere onto a surface of the octahedron inscribed inside the sphere. The faces of the octahedron then become the first 8 partitions in the hierarchy. The next level of the nested structure is recursively defined by 3 vertices and 3 bisects of each side of the triangle at any given level. Fig. 3 (after Kunszt¹⁶) illustrates the idea. The HTM software package also provides functions that return a set of triangles covering circular and polygonal areas. With proper indexing objects inside those triangles can be quickly searched through.

Although designed for slightly different applications, R-trees can also be adapted for spatial range queries of point objects, and have been reported to provide acceptable performance.¹⁸ R-tree based indexing is available in PostgreSQL and could be a viable alternative to HTM.

3. DATA MINING

The ultimate goal behind our efforts is to provide easy access to various variability data in order to enable extraction of new astronomical information for research on source variability. The web interface under construction will provide a set of high level tools for data analysis and visualization. Our strategy is to start from very basic functionality and gradually build a powerful data mining system. Ultimately, after extracting the required data, the user will have an option to run period searches, phase the data with an arbitrary period, perform other time series analysis, i.e., correlation functions, Fourier transforms and smoothing. The interface will allow users to plot various database stored and derived parameters, display sky maps with overlaid selection of objects, cross correlate various data sets and finally classify objects using both basic information as well as the information obtained in the course of the session. In the area of classification and clustering, Machine Learning has made remarkable progress in recent years. Given the growing complexity of astronomical data (high dimensionality with dimensions describing very diverse and sophisticated characteristics), mining large databases will require new efficient tools. The results from Machine Learning are typically more objective, since this approach often eliminates human bias and error.

Table 1. Results for variable star classification in Section 3.1.

Method		Accuracy	
		cross-validation data	training data
Supervised	SVM	90%	95%
	J4.8	90%	92%
	5-NN	81%	86%
Unsupervised	Autoclass	80%	80%
	k-Means	—	70%

3.1. Classification of Variable Stars

Currently we are evaluating machine learning algorithms for potential use in classification of astronomical objects. Two basic types of methods can be distinguished. Supervised techniques require a training sample with known class membership. Unsupervised learning, on the other hand, works on data with unknown classifications, including an unknown number of classes for some algorithms. Supervised methods naturally outperform unsupervised ones, however the use of training sample with known classifications imposes prior knowledge onto the final result and may not always be possible. Unsupervised algorithms are capable of finding entirely new classes of objects.

Using several supervised and unsupervised techniques we classified a set of ~ 1700 periodic variable stars from a pilot search of the ROTSE data.¹⁹ The data has been previously classified with the human made algorithm based on prior knowledge and the appearance of the scatter plots. In the study of the pilot ROTSE sample the labels have been verified by visual inspection of light curves. In Fig. 4 we show prototype objects for 8 classes out of 9 considered in our study (with the exception of the catch-all class “other”). The location of objects in parameter space is shown in Fig. 6. Classification is based on a light curve in a single photometric band only (period, amplitude, the ratio of first overtone to fundamental frequencies and the skewness of the magnitude distribution).

Table 1 summarizes our results for Support Vector Machines (SVM),²⁰ decision tree builder (J4.8), five nearest neighbors classifier (5-NN), k-means clusterer²¹ and Bayesian system Autoclass.²² More details on SVM tests on variable star data can be found in Ref. 23. Performance on training data is usually a poor predictor of the performance on future data. Much better results can be obtained from cross validation analysis. We performed a 5-fold cross validation, where randomly selected 4/5 of the sample is used for training and then the accuracy is evaluated on the remaining 1/5. With the state of the art SVM method, we achieved 90% accuracy for the full problem (9 classes). It was estimated that two human analysts would agree at a similar level. Actually for a fraction of cases, even the same person making the same classification twice, arrives at a different conclusion each time. Performance can be as good as 95 or even 98 % for a restricted problem when larger, more general classes are considered or when objects of one class are detected against the background of “everything else”. We have noted a very good potential for the use of decision trees. This type of algorithm is particularly attractive because it gives full insight into how the features were used to compute the classification, and the machine can then convey the knowledge back to a human. The tree in Fig. 5 correctly reproduces many of the features found in human made algorithm, despite the fact that we worked with a feature space somewhat different from the one used in the benchmark visual study. The tree was trimmed of the least populated branches and leaves to control over-fitting. We kept the final accuracy of the tree at 90%.

4. SUMMARY AND DISCUSSION

There are considerable scientific rewards coming out of massive data sets in astronomy. SkyDOT, a general sky variability database being designed and constructed at LANL, should have a major beneficial influence on the field. Several surveys either conducted locally or with indirect LANL involvement will provide the data for the database. The future synoptic data sets will be integrated into the database to maximize the scientific potential. The current activities focus on data conversion and web database application software. The first results in application of Machine Learning to this type of data are promising. Data Mining technologies will eventually be integrated with the database and may aid new surprising discoveries. The National Virtual Observatory bears the promise that astronomical community will be prepared for challenges in data overloaded astronomy. Recently the NVO, in collaboration with similar European initiatives, announced VOTable,⁵ a new XML based standard for exchanging astronomical data.⁶ We are closely following activities in this area to assure NVO compliant output from our database.

ACKNOWLEDGMENTS

This work is supported by the Laboratory Directed Research and Development funds at LANL under DOE contract W-7405-ENG-36.

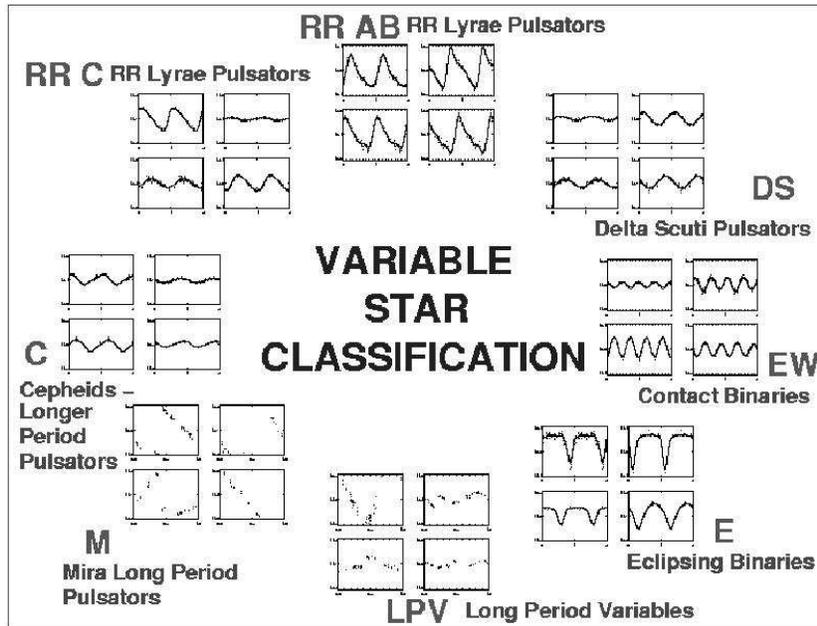


Figure 4. An illustration of variable star classification problem from Section 3.1. This example of classification is based entirely on light curve parameters (period, amplitude, and shape).

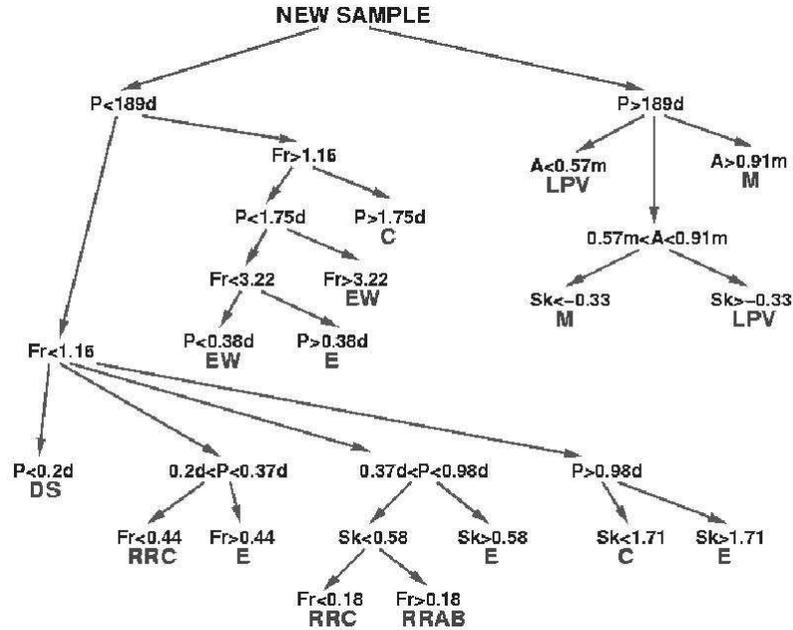


Figure 5. A binary tree obtained with the J4.8 algorithm for classification in Section 3.1. The main characteristics of the human classification are reproduced: e.g. period ranges at the bottom. The complexity of the tree was reduced using sub-tree rising by requiring a minimum 10 objects at any given leaf node and 90% confidence for pruning.

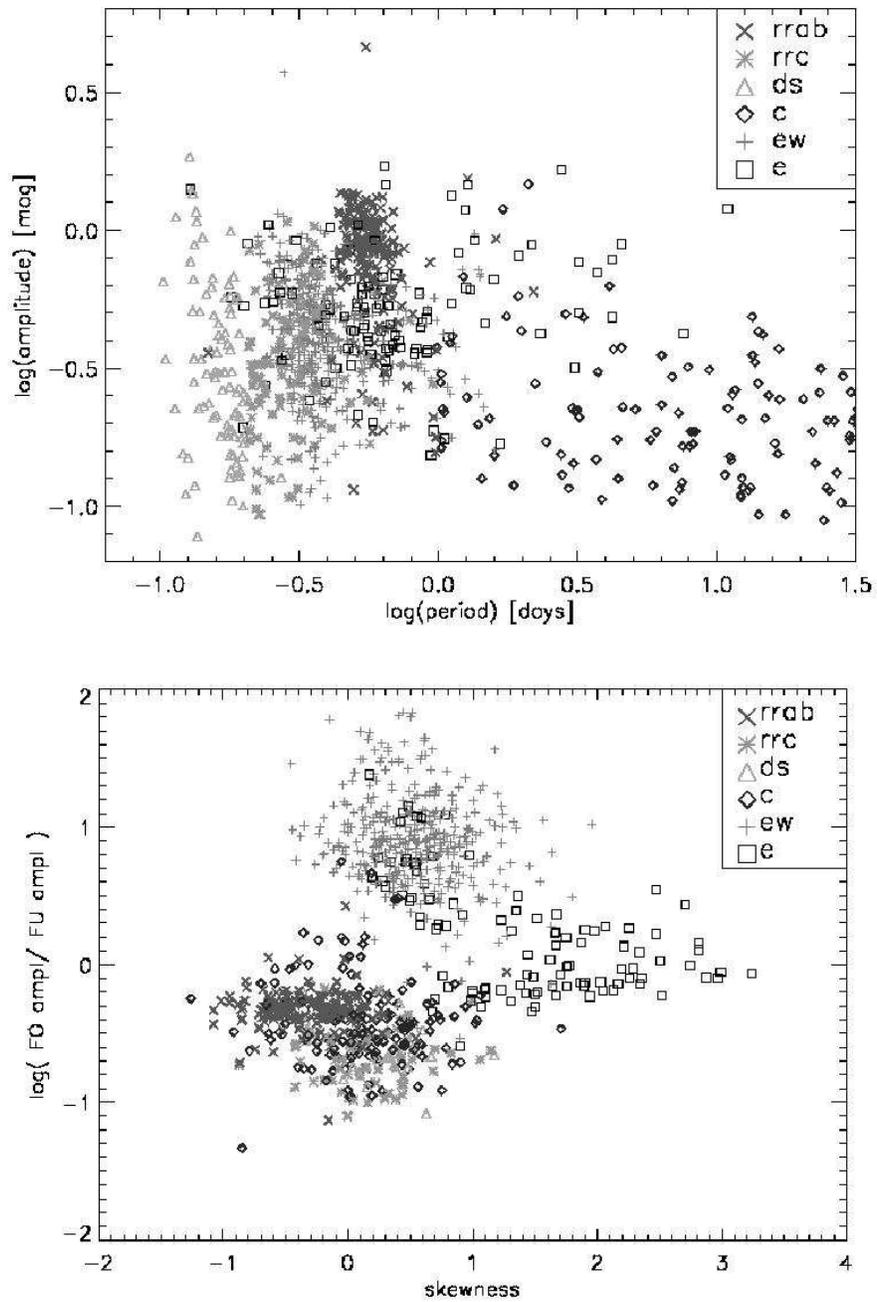


Figure 6. Feature space for our example of variable star classification. For a sample of 1700 periodic variables two projections are shown: Period-Amplitude (upper) and Skewness-Fourier ratio (lower).

REFERENCES

1. B. Paczynski, “Gravitational Microlensing in the Local Group”, *Ann. Rev. of Astron. & Astroph.* **34**, pp. 419–460, 1996.
2. B. Paczynski, “Current Status of the Microlensing Surveys”, in *The Impact of Large-Scale Surveys on Pulsating Star Research, ASP Conf. Proc.* **203**, pp. 9–18, 2000.
3. B. Paczynski, “The Future of Massive Variability Searches”, in *Variable Stars and the Astrophysical Returns of Microlensing Surveys*, R. Ferlet, J. P. Maillard and B. Raban, eds., *IAP Colloquium Proc.* **12**, pp. 357–371, 1997.
4. The US NVO White Paper, “Toward a National Virtual Observatory: Science Goals, Technical Challenges, and Implementation Plan”, in *Virtual Observatories of the Future*, R. J. Brunner, S. G. Djorgovski, and A. S. Szalay, eds., *ASP Conf. Proc.* **225**, pp. 353–357, 2001.
5. R. Williams, F. Ochsenbein, C. Davenhall, D. Durand, P. Fernique, D. Gaietta, R. Hanish, T. McGlynn, A. Szalay, and A. Wicenec, “VOTable: A proposed XML Format for Astronomical Tables”, <http://cdsweb.u-strasbg.fr/doc/VOTable/>, 2002.
6. F. Ochsenbein, M. Albrecht, A. Brighton, P. Fernique, D. Guillaume, R. Hanisch, E. Shaya, and A. Wicenec, “Using XML for Accessing Resources in Astronomy”, in *Astrophysical Data Analysis Software and Systems IX, ASP Conf. Proc.* **216**, pp. 83–87, 2000.
7. A. S. Szalay, J. Gray, A. R. Thakar, P. Z. Kunszt, T. Malik, J. Raddick, C. Stoughton, and J. vandenBerg, “The SDSS SkyServer: Public Access to the Sloan Digital Sky Server Data”, in *Management of Data, Proc. of ACM SIGMOD International Conf.*, in press, 2002.
8. R. J. Brunner, S. G. Djorgovski, T. A. Prince, and A. S. Szalay, “Massive Datasets in Astronomy”, in *Handbook of Massive Datasets*, J. Abello, P. M. Pardalos, and M. G. C. Resende, eds., Kluwer Academic Publishers, Dordrecht, 2002.
9. P. R. Wozniak, A. Udalski, M. Szymanski, M. Kubiak, G. Pietrzynski, I. Soszynski, and K. Zebrun, “Difference Image Analysis of the OGLE-II Bulge Data. III. Catalog of 200,000 Candidate Variable Stars”, *Acta Astronomica* **52**, pp. 129–142, 2002.
10. W. T. Vestrand, K. Borozdin, S. P. Brumby, D. Casperson, E. Fenimore, M. Galassi, G. Gisler, K. McGowan, S. Perkins, W. Priedhorsky, D. Starr, R. White, P. Wozniak, and J. Wren, “Searching for Optical Transients in Real-Time. The RAPTOR Experiment.”, in *Gamma-Ray Bursts and Afterglow Astronomy*, G. Ricker, ed., *Proc. of the Woodshole GRB Workshop*, in press, 2001.
11. C. Akerlof, R. Balzano, S. Barthelmy, J. Bloch, P. Butterworth, D. Casperson, T. Cline, S. Fletcher, F. Frontera, G. Gisler, J. Heise, J. Hills, K. Hurley, R. Kehoe, B. Lee, S. Marshall, T. McKay, A. Pawl, L. Piro, J. Szymanski, and J. Wren, “Prompt Optical Observations of Gamma-Ray Bursts”, *Astroph. J. Letters* **532**, pp. 25–29, 2000.
12. A. Udalski, M. Kubiak, and M. Szymanski, “Optical Gravitational Lensing Experiment. OGLE-2 – the Second Phase of the OGLE Project”, *Acta Astronomica* **47**, pp. 319–344, 1997.
13. W. T. Vestrand, K. N. Borozdin, S. P. Brumby, D. E. Casperson, E. E. Fenimore, M. C. Galassi, K. McGowan, S. J. Perkins, W. C. Priedhorsky, D. Starr, R. White, P. Wozniak, and J. Wren, “RAPTOR experiment: a system for monitoring the optical sky in real time”, in *Advanced Global Communications Technologies for Astronomy II*, J. McConnell, ed., *Proc. SPIE* **4845**, in press, 2002.
14. P. R. Wozniak, C. Akerlof, D. Casperson, G. Gisler, R. Kehoe, B. Lee, S. Marshall, K. E. McGowan, T. McKay, E. Rykoff, D. A. Smith, W. T. Vestrand, and J. Wren, ROTSE Collaboration, “An All-Sky Variability Census using ROTSE-I”, *Bull. Am. Astron. Soc.* **200**, 58.02, 2002.
15. T. M. Connolly, C. E. Begg, and C. Begg, *Database Systems: A Practical Approach to Design, Implementation, and Management*, Addison-Wesley Publishing, Boston, 2001.
16. P. Z. Kunszt, A. S. Szalay, I. Csabai, and A. R. Thakar, “The Indexing of the SDSS Science Archive”, in *Astronomical Data Analysis Software and Systems IX*, N. Manset, C. Veillet, and D. Crabtree, eds., *ASP Conf. Proc.* **216**, pp. 141–145, 2000.
17. W. O’Mullane, A. J. Banday, K. M. Górski, P. Kunszt, and A. S. Szalay, “Splitting the Sky - HTM and HEALPix”, in *Mining the Sky*, A. J. Banday, S. Zaroubi, and M. Bartelmann, eds., *Proc. of the MPA/ESO/MPE Workshop in Garching*, pp. 638–642, 2001.

18. A. Baruffolo, “R-Trees for Astronomical Data Indexing”, in *Astronomical Data Analysis Software and Systems VIII*, D. M. Mehringer, R. L. Plante, and D. A. Roberts, eds., *ASP Conf. Proc.* **172**, pp. 375–378, 1999.
19. C. Akerlof, S. Amrose, R. Balsano, J. Bloch, D. Casperson, S. Fletcher, G. Gisler, J. Hills, R. Kehoe, B. Lee, S. Marshall, T. McKay, A. Pawl, J. Schaefer, J. Szymanski, and J. Wren, “ROTSE All-Sky Surveys for Variable Stars. I. Test Fields”, *Astronom. J.* **119**, pp. 1901–1913, 2000.
20. V. Vapnik, *Statistical Learning Theory*, Wiley-Interscience, New York, 1998.
21. I. H. Witten, and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann Publishers, San Francisco, 1999.
22. P. Cheeseman, and J. Stutz, “Bayesian classification (AutoClass): Theory and results”, in *Advances in Knowledge Discovery and Data Mining*, U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, eds., pp. 153–181, AAAI Press/MIT Press, Cambridge, Mass., 1996.
23. P. R. Wozniak, C. Akerlof, S. Amrose, S. Brumby, D. Casperson, G. Gisler, R. Kehoe, B. Lee, S. Marshall, K. E. McGowan, T. McKay, S. Perkins, W. Priedhorsky, E. Rykoff, D. A. Smith, J. Theiler, W. T. Vestrand, and J. Wren, ROTSE Collaboration, “Classification of ROTSE Variable Stars using Machine Learning”, *Bull. Am. Astron. Soc.* **199**, 130.04, 2002.