New Horizons in Time Domain Astronomy
Proceedings IAU Symposium No. 285, 2011
R.E.M. Griffin, R.J. Hanish & R. Seaman, eds.
© 2011 International Astronomical Union
DOI: 00.0000/X000000000000000X

Classification of ASKAP VAST Radio Light Curves

Umaa Rebbapragada[1] & Kitty Lo[2] & Kiri L. Wagstaff[1] & Colorado Reed[3] & Tara Murphy[2] & David R. Thompson[1]

[1]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, 91109 USA
email: Umaa.D.Rebbapragada@jpl.nasa.gov

[2]Sydney Institute for Astrophysics, School of Physics, University of Sydney, Sydney, NSW Australia

[3]Department of Physics, University of Iowa, Iowa City, IA USA

Abstract. The VAST survey is a wide-field survey that observes with unprecedented instrument sensitivity (0.5 mJy or lower) and repeat cadence (a goal of 5 seconds) that will enable novel scientific discoveries related to known and unknown classes of radio transients and variables. Given the unprecedented observing characteristics of VAST, it is important to estimate source classification performance, and determine best practices prior to the launch of ASKAP's BETA in 2012. The goal of this study is to identify light curve characterization and classification algorithms that are best suited for archival VAST light curve classification. We perform our experiments on light curve simulations of eight source types and achieve best case performance of approximately 90% accuracy. We note that classification performance is most influenced by light curve characterization rather than classifier algorithm.

Keywords. ASKAP, VAST, radio astronomy, classification, data analysis

1. Introduction

The Australian Square Kilometer Array Pathfinder (ASKAP) will observe the entire visible radio sky, including previously unexplored regions of phase space, in a single day with sub-mJy sensitivity at a 5-second cadence. Because no other telescope in operation has these capabilities, ASKAP has the potential to significantly advance the study of known transients and variables, while enabling the discovery of new objects and object classes. The Variables and Slow Transits (VAST) survey science project of ASKAP is focused on the development of new algorithms for the detection of transients with timescales as short as 5 seconds (Murphy & Chatterjee (2009)). Source types of interest include X-Ray Binaries, Supernovae, Extreme Scattering Events, Intra-Day Variables, Novae, and Flare Stars dMe and RSCVn. Source classification is a prerequisite for scientific study of radio transients and variables.

The overall goal of our study is to evaluate state-of-the-art machine learning methods for archival classification of VAST light curves. Because VAST has no existing counterpart with data to use for empirical evaluation, we simulate light curves for each of the sources discussed above plus background sources, and study performance using different classifiers, observing strategies, signal-to-noise ratios, and light curve characterizations.

For space reasons, we present summary results using a daily observational strategy (VAST Wide (Murphy & Chatterjee (2009))) that observes with an RMS of 0.5mJy. Our results show that we achieve approximately 90% classification accuracy using a Support Vector Machine and a concatenation of different feature representations. These results and others will be published as a VAST Memo in early 2012.

2. Classifiers and Light Curve Characterizations

We selected standard classification algorithms that have proven successful on other light curve classification tasks (Richards *et al.* (2011) and Wachman *et al.* (2009)). Specifically, we evaluated Support Vector Machine (SVM (Cortes & Vapnik (1995))), Decision Tree (J48 (Quinlan (1986))), and Random Forest (Breiman (2001)) classifiers using implementations provided by the Weka data mining package (Hall *et al.* (2009)). We have also worked with other types of classifiers, including probabilistic classifiers such as Naive Bayes and Logistic Regression. However, we found that SVMs, Decision Trees and Random Forests produced superior results on this data.

Machine learning methods for classification presume the existence of a structured data set, where each example is a vector of "features." Real light curve observations may not meet this requirement because light curves may contain different numbers of observations taken at differing sampling rates. Thus, in a real setting, one must create representations of the data that meet these requirements.

Our first feature set extracts statistics from the flux measurements of each light curve. These are a subset of the "non-periodic statistical features" used by Richards *et al.* (2011) and include moment statistics (e.g., mean, standard deviation, skew, kurtosis), flux percentile ratios, and shape statistics. We refer to this feature set as *stat*. Our second feature set, *lsp*, extracts coefficients from the Lomb-Scargle Periodogram representation of each light curve (Scargle (1982)). Specifically, we extract power information from the top 20 frequencies. Our third feature set, *wlet*, extracts wavelet coefficients using the discrete wavelet transform (DWT).

From the original time domain observations, the statistical and two frequency space characterizations, we create the following six feature sets for our experiments: tme (time domain flux measurements), stat, lsp, wlet, all-reps (concatenation of stat, lsp, and wlet), and all (concatenation of tme, stat, lsp and wlet).

3. Experimental Setup and Results

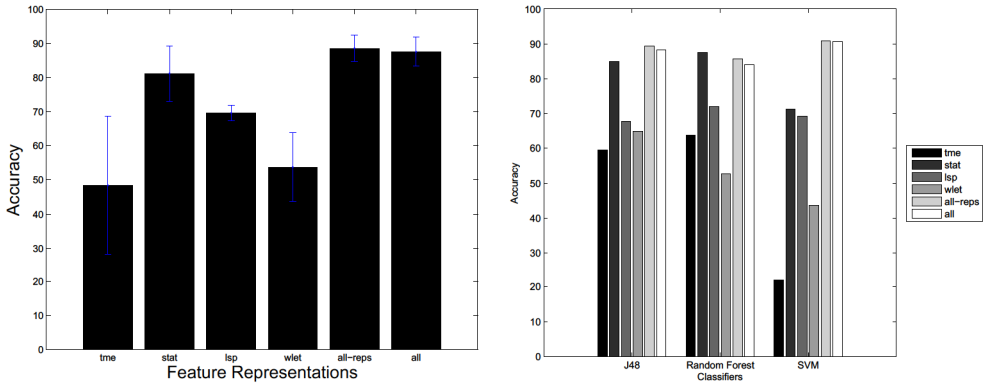
We simulated 200 400-day light curves per source type at signal-to-noise ratios of 3, 5, 7 and 10 (each is a unit of standard deviation). For source types Intra-day Variables and Extreme Scattering Events, SNR is defined in relation to the source's quiescent flux. For all other (transient) source types, SNR is defined with respect to the source's peak flux. For transient source types, the event occurs at time 0.

Our first result in Figure 1(a) shows accuracy by feature representation, averaged across all other parameters (SNR and classifier). We measure accuracy using 10-fold cross validation. The results show that the time domain observations alone yield the weakest performance on average, and combining the feature representations (all and all-reps) yields the best performance.

Figure 1(b) shows accuracy per classifier and feature, averaged across SNR. SVM seems to have the largest variability in performance, recording the lowest performance for the tme feature, but the highest performance for the all-reps feature set. For the higher-performing all-reps and all feature representations, the three classifiers perform similarly. We conclude that feature representation more strongly informs performance than classifier selection.

4. Conclusions

These results are part of ongoing work to estimate classification perform of VAST data prior to the arrival of commissioning data from ASKAP's BETA. We have also



(a) Accuracy by feature representation

(b) Accuracy by classifier and feature representation

Figure 1. Classification accuracy

studied the impact of different VAST observational strategies, and estimated classification performance per source type. We plan to publish those results along with the results in this paper in a VAST Memo in early 2012. Our future plans are to refine our methods and feature representations in order to optimize classification performance in the archival setting.

5. Acknowledgements

This work was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration. Copyright 2011. Government sponsorship acknowledged.

References

Breiman, L. 2001, *Machine Learning* 45, p. 5

Cortes, C. and Vapnik, V. 1995, *Machine Learning* 20, p. 273

Hall, M. , Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H. 2009, *SIGKDD Explorations* 11, 1

Quinlan, J. R. 1986, *Machine Learning* 1, 1

Richards, J. W., Starr, D. L., Butler, N. R., Bloom, J. S., Brewer, J. M., Crellin-Quick, A., Higgins, J., Kennedy, R., Rischard, M. 2011, *ApJ* 733, 1

Scargle, J. D. 1982 *ApJ* 263, p. 835

Wachman, G., Khardon, R., Protopapas, P., Alcock, C. R. 2009, in: W.L. Buntine, M. Grobelnik, D. Mladenic & J. Shawe-Taylor (eds.), *Kernels for periodic time series arising in astronomy*, Proc. ECML (Bled, Slovenia), p. 489

Murphy, T., Chatterjee, S. 2009, http://www.physics.usyd.edu.au/sifa/vast/index.php/Main/Documents