Memo 5

Offline and Online Classification of Simulated VAST Transients

Umaa Rebbapragada et al.

February 29 2012

VAST Memo
Offline and Online Classification of Simulated VAST Transients

Umaa Rebbapragada[1], Kitty Lo[2], Kiri L. Wagstaff[1], Colorado Reed[3], and Tara Murphy[2]

[1]Machine Learning and Instrument Autonomy, California Institute of Technology, Jet Propulsion Laboratory, Pasadena, CA USA
[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

February 29, 2012

Abstract

VAST is an unprecedented wide-field survey planned with ASKAP, the Australian SKA Pathfinder, that will enable novel scientific discoveries related to known and unknown classes of radio transients and variables. The VAST data processing pipeline extracts sources from 5-second images and builds light curves that are stored in a data archive for science user consumption. This memo addresses two source classification tasks that occur within the pipeline. The first is at the archive level where science users may issue queries for known source types (offline classification). The second occurs during real-time processing in order to trigger appropriate follow up when transient phenomena are detected (online classification). Both tasks require automated methods to classify sources in the time domain. Given the unprecedented observing characteristics of VAST, it is important to estimate classification performance in both settings, and determine best practices prior to the commissioning of ASKAP's BETA in 2012. This memo identifies candidate light curve characterizations and classification algorithms, and studies their performance under different observing strategies and levels of noise in both offline and online settings. Our results show that the choice of light curve characterization influences classification performance more than the selection of learning algorithm, and that a combination of feature sets yields best performance. We achieve approximately 93% and 70% classification accuracy in the offline and online cases respectively. Classes that are commonly confused include novae versus supernovae and ESEs versus background sources.

1 Introduction

The Australian Square Kilometer Array Pathfinder (ASKAP) will allow us to probe unexplored regions of phase space. In a single day, ASKAP can scan the entire visible sky and achieve sensitivity of 0.5 mJy or lower. 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.

VAST, an ASKAP Survey for Slow Variables and Transients, is focused on the detection of variables and transients with timescales as short as 5 seconds. Table 1 lists source types of particular

Source Type	Abbreviation	Simulated
Gamma Ray Bursts	GRB	
X-Ray Binaries	XRB	✓
Supernovae	SNe	✓
Intra-Day Variables	IDV	✓
Extreme Scattering Events	ESE	✓
Novae	Novae	✓
Flare Stars dMe	dMe	✓
Flare Stars RSCVn	RSCVn	✓
Cataclysmic Variables	CV	
SGR-like flares	SGR	

Table 1: Some of the variables and transients of scientific interest to the VAST survey. The "Simulated" column indicates the source types included in this study.

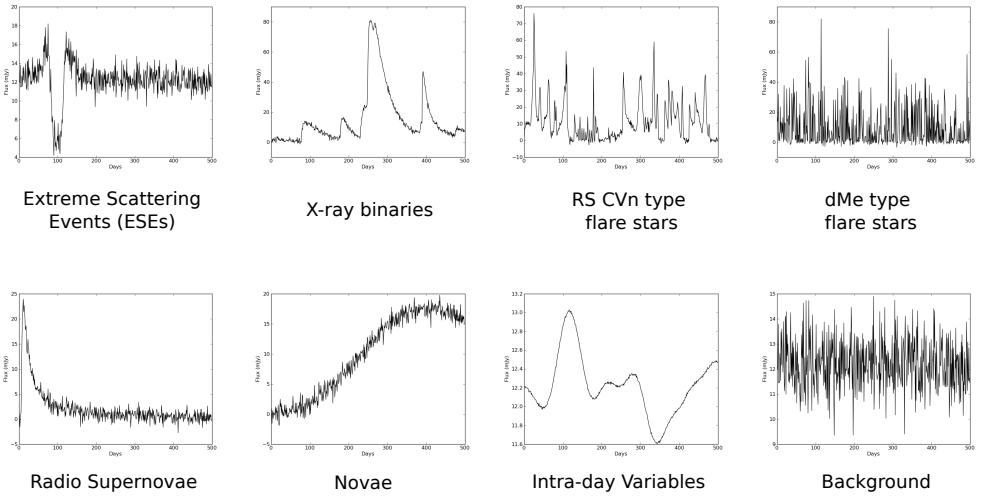


Figure 1: Simulated lightcurves of eight source types.

interest to VAST and Figure 1 shows some example simulated lightcurves of each source type. VAST Memo #1 (Murphy et al. [2009]) details the scientific case for investigating each source type and the advances that VAST may enable. For example, VAST allows for the first time a comparison of radio and optically-observed core-collapse supernovae, and possibly new supernovae that were previously undetected at infrared and optical wavelengths due to intervening dust (Murphy et al. [2009]). Because VAST observes a wide field every day, it is well-suited for detecting sub-day variable IDVs, and may help determine a physical model for ESEs (Murphy et al. [2009]). Of course, the most exciting prospect of VAST is its potential to discover new objects and object classes.

Because source classification is a prerequisite for scientific study of radio transients and variables, it has a pivotal role in the VAST data processing pipeline, depicted in Figure 2. The yellow shaded boxes indicate the locations at which source type classification is performed: at the archive level (offline classification) and during real-time processing following transient detection and light curve extraction (online classification). Offline classification of the archive enables targeted queries by

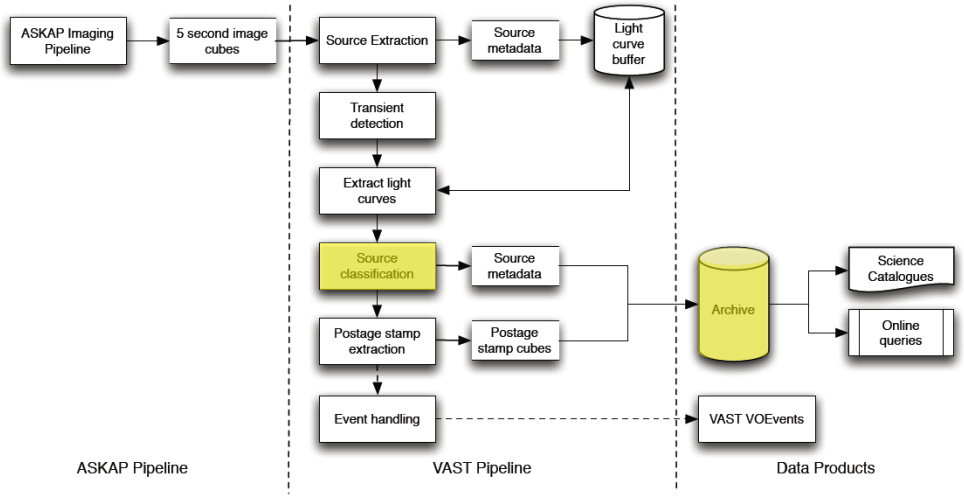


Figure 5: The VAST transient detection pipeline.

Figure 2: VAST data processing pipeline (Murphy et al. [2009], Banyer et al. [2012]) with source classification highlighted.

end users for specific source types. Because this classification task is performed offline, it may accept a heavier computational cost for the incorporation of additional input variables, the use of a more robust classifier, or lookups in other catalogs.

Online classification of newly-detected transient sources in real-time enables a tailored follow-up response depending upon the source type classification. For example, if the detected event is classified as a GRB or SNe, then follow-up at optical wavelengths is desirable, while ESE classification may trigger observations at VLBI wavelengths (Murphy et al. [2009]). Online classification is a more challenging problem because of the computational and timing constraints of real-time processing. Decisions must be made after only a few observations are recorded.

This study's goal is to evaluate state-of-the-art machine learning methods for offline and online classification of VAST light curves. Because VAST has no existing counterpart with data to use for empirical evaluation, we constructed simulated light curves for the source types indicated in Table 1. We also simulated different observing strategies, signal-to-noise ratios, and light curve characterizations in order to assess their impact on classification performance. Our simulations consist of seven transient and variable source types plus a background source type (BG), using six VAST observing strategies (Murphy et al. [2009]).

We present overall classification results and results per source type in both the offline and online settings. Our offline results explore the impact of classifier selection, light curve characterization, and increasing levels of noise. We find that light curve characterization more strongly influences classification performance than learning algorithm selection, and that the concatenation of all

feature representations yields best performance. In the online case, we evaluate three different online classification methods using two learning algorithms and two light curve characterizations. We find the best performance from models that were trained on historical light curves truncated to the same length as the current observation, as opposed to training on the full history of observations. Also, feature sets built cumulatively are more informative than statistics calculated once from all available observations. We achieve $\sim 93\%$ and $\sim 50\%$ accuracy averaging across all source types in the offline and online cases respectively. In the online case, we show that certain source type groupings increase classification performance to $\sim 70\%$.

2 Simulated VAST Data

We generated several sets of 400-day light curves for source types XRBs, SNe, IDVs, ESEs, Novae, dMe, RSCVn and BG, in which we varied the observational strategy and signal-to-noise ratio[1]. All transient events begin at time 0. Those that do not last 400 days are padded with white noise.

The observational strategies that we simulated are:

- **VAST Wide (WIDE)**: Sampled once per day with sensitivity (rms) 0.5 mJy.
- **VAST Galactic Plane (GP)**: Irregularly sampled at least once per week and at most once daily, with rms 0.1 mJy.
- **VAST Deep (DEEP)**: Sampled once per day on days 1, 2, 3, 4, 17, and 21 at rms 50 μJy.
- **logarithmic (LOG)**: Sampled once per day on days 1, 2, 4, 8, etc. at rms 0.5 mJy.
- **monthly (MONTHLY)**: Sampled once every 30 days at rms 0.5 mJy.
- **patches (PATCHES)**: Three consecutive days randomly chosen per month at rms 0.5 mJy.

WIDE, GP, and DEEP are identified in VAST Memo #1 (Murphy et al. [2009]) as candidate observing strategies for VAST. LOG, MONTHLY, and PATCHES were conceived as possible alternatives.

In addition to simulating light curves from several observational strategies, we simulated data at four signal-to-noise ratio (SNR) levels: 3.0, 5.0, 7.0 and 10.0 (in units of standard deviation). In the case of variable source types (IDVs and ESEs), SNR is defined in relation to the source's mean flux. For all other (transient) source types, SNR is defined with respect to the source's peak flux. For transient source types, the event begins at time 0. Table 2 provides a summary of our simulation parameters. We generated 200 light curves for each source type, survey strategy, and SNR value combination listed.

2.1 Known Limitations of the Simulated Data

Our simulations do not currently model instrumentation problems, effects of bad weather, and red noise. As such, the simulated light curves contain more observations and less noise than would be expected in real VAST data. Furthermore, we used balanced training and test sets containing an equal number of sources per class. Thus, the results from this study are an upper bound on expected classification performance from real VAST data given the current set of techniques and algorithms.

[1]The light curve simulation process will be documented in a forthcoming VAST Memo.

Source Types	Survey Strategies (Number of Observations per Light Curve)	SNR
XRBs, SNe, IDVs, ESEs, Novae, dMe, RSCVn, and BG	WIDE (400), DEEP (6), GP (100), LOG (9), MONTHLY (14), PATCHES (39)	3σ, 5σ, 7σ, 10σ

Table 2: Summary of simulated data parameters. All sources were simulated for 400 days.

3 Classification Tasks

We anticipate offline and online classification will be integral parts of the VAST data processing pipeline (see Figure 2). Offline classification will be performed on a data archive containing source light curves and meta-data. It may also use more computationally expensive classification methods than online classification, and can in theory benefit from enhanced feature sets gained through accessing other catalogs with complementary observations at other wavelengths. In this study, however, we do not explore the benefits of such additional information. Because the sources in this study were simulated, they have no real counterparts in other catalogs to provide supplementation observations. Therefore, we evaluated classifier performance using only the 400 days of simulated light curve observations.

Online classification will be conducted as soon as new observations arrive. Newly detected transients will initially be represented by a single observation that is augmented as time goes on. For such sources, the earlier a classification decision is made, the more likely that a follow-up observation by a complementary instrument can be performed. Transient science relies on well coordinated early-time multi-wavelength follow-up observations. Therefore, the system will often be called upon to make classification decisions using only a few observations. We evaluated online classification performance using the first 1 to 30 days of simulated observations.

4 Light Curve Representations

Machine learning methods for classification presume the existence of a structured data set, where each light curve can be represented as "feature vectors" of identical length. Raw light curve observations may not meet this requirement due to differing sampling rates and missing observations. One must create homogenized light curve representations. We worked with the following three representations.

4.1 Non-Periodic Statistics

We extracted 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]. We group them as follows:

- **Moment statistics**: *mean*, *standard deviation*, *skew*, and *kurtosis*. We also calculated *beyond1_std*, which is the fraction of magnitudes that are either above or below one standard deviation.

- **Flux percentile ratios**: We define a flux percentile $F_{n,m}$ to be the difference between the flux values at percentiles n and m, and use the following flux percentile ratios:
  - *mid-20*: $F_{40,60}/F_{5,95}$
  - *mid-35*: $F_{32.5,67.5}/F_{5,95}$
  - *mid-50*: $F_{25,75}/F_{5,95}$
  - *mid-65*: $F_{17.5,82.5}/F_{5,95}$
  - *mid-80*: $F_{10,90}/F_{5,95}$
  - *percent different flux*: $F_{2,98}$
  - *percent amplitude*: Largest percentage of deviation from median flux.
- **Morphological/Miscellaneous**: The following are calculated from flux magnitudes as follows:
  - *max slope*: Maximum slope value between two adjacent observations.
  - *amplitude*: Difference of the maximum and minimum flux measurements.
  - *median absolute deviation*: Median of deviations from the median value: $median(|mag - median(mag)|)$.
  - *median buffer range percentile*: Fraction of observations within 20% of median magnitude.
  - *positive slope trend*: Fraction of adjacent observations with positive slope.
  - *modulation index*: RMS divided by the average flux.

4.2 Wavelet Coefficients

The wavelet representation of a time series contains both temporal and frequency information. We use the simplest wavelet basis, the Haar wavelet, and extract the wavelet coefficients using the discrete wavelet transform (DWT). The DWT takes a time series, with sampling frequncy of f, and passes it through a high pass and low pass filter such that they consist of information from frequencies $f/2$ to f and 0 to $f/2$ respectively. Then, the wavelet coefficients are derived by multiplying and then summing the wavelet basis function with the output of the high pass filter. For the Haar wavelet, this means differencing the adjacent values. The output of the low pass filter is put through the band pass filter step again to derive the next level of wavelet coefficients which sample the lower frequency space. This is iterated until the lowest frequency level is reached.

The DWT requires the time series to have 2^n time points. If this condition is not fulfilled, then the time series need to be padded with extra time points. We use the python package PyWavelets to extract the wavelet coefficients. The default padding option in PyWavelet is to fill the time series with symmetric entries.

4.3 Lomb-Scargle Periodogram

We use the Lomb-Scargle Periodogram (LSP) as a frequency space representation of unevenly-sampled light curves (Scargle [1982]). The LSP is a "slight modification" to the classical periodigram, which calculates the discrete Fourier Transform (DFT) of a signal and outputs the power of

Survey Strategies	Batch (k) Size (# Observations)	Num Batches
WIDE	5	8
GP	20	5
PATCHES	3	13
MONTHLY	6	2
LOG	9	1
DEEP	6	1

Table 3: Process by which the stat-cum representation is constructed. Statistical features are computed for each batch of observations.

a given frequency. The LSP is equivalent to a least-squares fit of sine waves to the data, and is also equivalent to the classifical periodigram for evenly-spaced data. However, evenly-spaced data determine a natural set of frequencies (e.g., $\omega_n = 2\pi dt$ where dt is the length of time between adjacent points) and a well-defined Nyquist frequency. In order to determine an appropriate frequency range for the LSP, we follow the example of Scargle [1982] and oversample the "natural frequency" range by a factor of two. This often produces identical power information when tested on evenly sampled data. Once the LSP is calculated, we extract power information from the top 20 frequencies for VAST WIDE, GP, PATCHES, and MONTHLY. For the LOG and DEEP observing strategies with 9 and 6 observations, we use all 18 and 12 frequencies respectively.

4.4 Feature Sets

Using the original time domain observations, plus the statistical and frequency space characterizations, we create the following seven feature sets for our experiments:

- **Time Domain (tme)**: Time domain flux measurements.
- **Statistical (stat)**: The 18 non-periodic statistical features discussed in Section 4.1.
- **Cumulative Statistics (stat-cum)**: Statistical (stat) features built cumulatively after observing a fixed number of observations in multiples of k (k, $2k$, etc.). Table 3 describes how stat features are cumulatively built for each observing strategy. Strategies DEEP and WIDE contain too few observations to subdivide. Thus, the stat-cum and stat feature sets are equivalent for those strategies.
- **Lomb-Scargle Periodogram (lsp)**: The top (largest) 12, 18 and 20 frequencies for DEEP, LOG and all other survey strategies, respectively (as discussed in Section 4.3).
- **Wavelet (wlet)**: The approximation coefficient and the details coefficients at all possible levels. We used the Python module PyWavelets and the Haar wavelet.
- **All Feature Representations (all-reps)**: The concatenation of the stat-cum, lsp, and wlet feature sets.
- **All (all)**: The concatenation of all feature sets (tme, stat-cum, lsp, and wlet).

5 Classification Methods

5.1 Offline Classification

We selected standard classification algorithms that have proven successful on other light curve classification tasks (Richards et al. [2011], Wachman et al. [2009]). Specifically, we evaluated support vector machine (SVM) (Cortes and Vapnik [1995]), decision tree (J48) (Quinlan [1986, 1993]), and Random Forest (RF) (Breiman [2001]) classifiers, using the 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.

The SVM is a binary classifier that calculates a decision boundary that maximizes the distance between training examples and a separating hyperplane (Bishop [2006]). SVMs are powerful and flexible decision machines because they make use of kernel substitution for transforming a linearly inseparable input space into a higher-dimensional space that is linearly separable. SVMs extend to multi-class classification problems through the *one-versus-one* approach where $\binom{K}{2}$ SVMs are trained separately to discriminate between each possible pair of the K classes, and then the results of all classifiers are combined to classify a new observation. Weka's SVM implementation uses the sequential minimal optimization (SMO) (Platt [1999], Keerthi et al. [2001]) method to train the SVM. We specified a radial basis function (RBF) kernel ($C = 1$, $\gamma = 0.1$).

A decision tree is a combination of simple decision models constructed on different regions of input space (Quinlan [1986, 1993], Bishop [2006]). The classifier forms a tree model where the pathway from root to leaf forms a conjunction of input variable tests with a class label. Whether an input proceeds down the right or left branch of a node is based on a positive or negative answer to the question "Is feature $f_i < c$?" for some c within the range of values for f_i. After the decision tree is built, some nodes or sub-trees are "pruned" and replaced with leaves in order to reduce model complexity and prevent overfitting. A Random Forest (Breiman [2001]) grows multiple unpruned decision trees by sampling the training set with replacement. The final decision is made by combining the predictions of individual trees. The advantage of using tree-based models like decision trees and Random Forests is that they are fast to build, generate human-interpretable models, and serve as indicators of feature relevance (by position in the tree). We used Weka's implementation of the C4.5 decision tree algorithm (Quinlan [1993]), called J48, and Weka's Random Forest implementation that builds a 10-tree ensemble.

5.2 Online Classification

In the online setting, we restricted our experiments to the critical first 30 days of observations of each source. Thirty days is an arbitrary time that we have chosen for this online experiment and the actual timeliness requirement depends on the source type and other factors.

Given a new source to classify that has been observed for t days, we considered the following strategies:

- 400-day-fixed: Train a classifier using features extracted from archival 400-day light-curves, then classify the new source using all available observations (up to 30 days).
- t-day-fixed: Train a classifier on all previously seen sources using only their observations from day 1 to day t, then classify the new source observed with t days of observations.

- t-day-adaptive: Train an ensemble of classifiers for all possible observing durations (here, for $t = 1$ to 400). Starting with the classifier for $t = 1$, classify the new observation using its observations from only day 1. Refrain from committing to a classification unless the posterior confidence from the classifier exceeds a desired threshold. If not, continue collecting observations of the new source and increasing t, employing the corresponding trained classifier, until the desired confidence is reached. This flexible approach enables a faster decision for easily classified sources, rather than waiting for the same fixed number of t days for all sources (desJardins et al. [2010]).

We limited the lightcurve feature representation to only the non-periodic statistical feature sets (stat and stat-cum). The wavelet (wlet) and Lomb-Scargle periodogram (lsp) coefficients are not meaningful with so few samples. We implemented our experiments with the J48 and Random Forest classifiers.

6 Results

6.1 Offline Classification

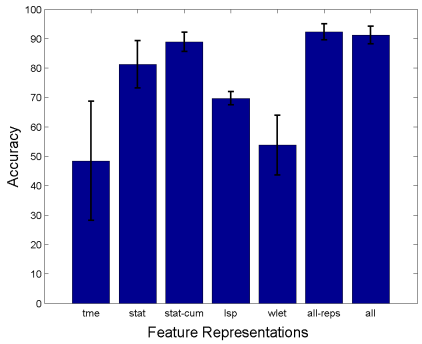
For each source type, we generated a master data set of 200 light curves each containing 400 observations sampled once daily. With eight source types, the data set contained 1600 light curves. We applied each observational strategy to the master data set, essentially removing observations not part of the strategy. Note that WIDE, which is sampled once daily, uses this data set in its original form unchanged. We measure the accuracy of offline classification performance using 10-fold cross validation. We also look at accuracy per source type and examine class confusions.

Figure 3 shows a series of results that show the effect of different feature representations, classifiers, SNRs, and observational strategies on accuracy using WIDE. Our first result in Figure 3(a) shows accuracy by feature representation, averaged across all other parameters (SNR and classifier). The results show that the use of time domain observations alone yields the weakest performance on average, and combining the feature representations (all and all-reps) yields the best performance.

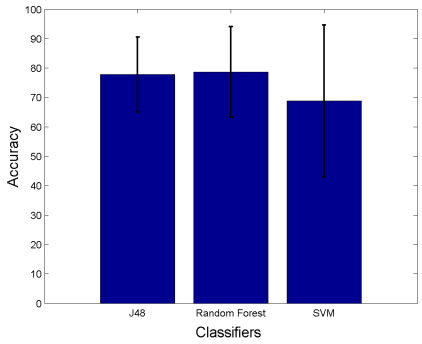
Figure 3(b) shows accuracy per classifier, averaged across SNR and feature representation. The results show that Random Forest has a slight edge over J48, followed by SVM. SVM seems to have the largest variability in performance, a fact that is confirmed in Figure 3(c). SVM records the lowest performance 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.

Figure 3(d) shows for all-reps and all that an increase in SNR improves classification accuracy. However, we were expecting more dramatic improvement from an SNR increase of 3.0 to 10.0. We suspect that our results are influenced by our method of simulation where both rms and SNR are held fixed, and the strength of the signal varies to achieve the desired SNR.

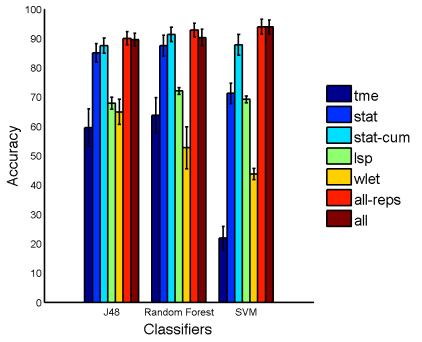
Figure 3(e) shows the effect of observational strategy on accuracy results using feature representations all-reps and all at SNR 5.0. The observational strategies are in descending order of number of light curve observations. Thus, we see performance decrease towards the right side of the plot. WIDE and GP have comparable performance, even though GP has one quarter the observations that WIDE does (100 versus 400). PATCHES has 40% of the observations of GP, yet achieves ∼80%



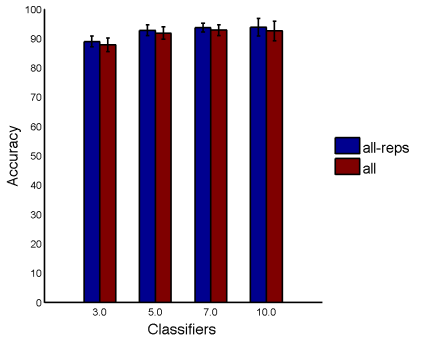
(a) Accuracy by feature representations



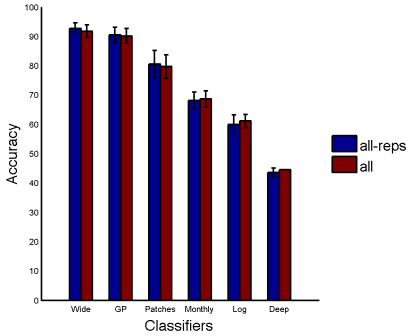
(b) Accuracy by classifier



(c) Accuracy by classifier and feature



(d) Accuracy by SNR and feature



(e) Accuracy by Observing strategy and feature

Figure 3: Offline classification results

accuracy versus GP's 88%. PATCHES, MONTHLY, LOG and DEEP have 39, 14, 9, and 6 observations and average accuracies of 80%, 67%, 60% and 43%. Note that 3 additional observations yields a 17% increase in accuracy, yet 5 additional observations yield only a 7% improvement. Early observations dramatically affect classification performance.

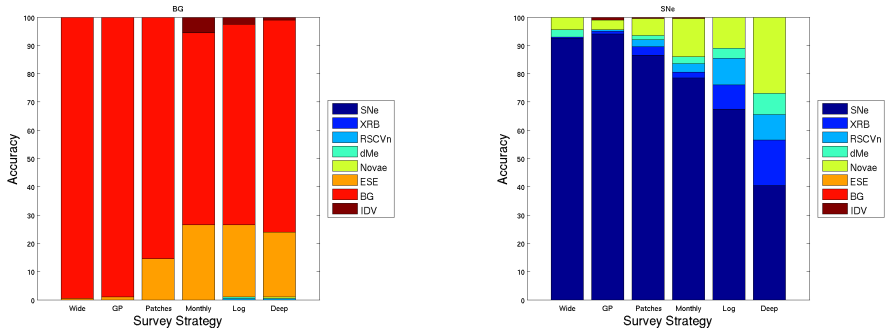
6.1.1 Accuracy Per Source Type

We now analyze accuracy per source type and observational strategy for classifier SVM and feature set all-reps. Figure 4 depicts class confusions. Key observations are as follows:

- **BG**: Background sources are almost perfectly classified with WIDE and GP, and over 85% for PATCHES. With 14 or fewer observations (MONTHLY, LOG and DEEP), classification accuracy dips below 80%, with confusion primarily with ESEs.
- **SNe**: WIDE and GP have at least $\sim$ 90% accuracy. The primary confusion is with Novae. When the number of observations drops to 14, accuracy decreases to less than 80%, with confusion increasing to include both flare star types and XRBs.
- **RSCVn**: WIDE has least 85% accuracy with primary confusion with dMe and XRB. Classification performance decreases dramatically starting with MONTHLY (less than 50%). Confusion with Novae increases with LOG.
- **dMe**: WIDE and GP have 100% and at least 95% accuracy. Primary confusers are RSCVn, and XRB and SNe starting with MONTHLY, and Novae starting with LOG.
- **Novae**: Novae are classified with greater than 90% accuracy at observations of 9 (LOG) and higher, with small levels of confusion with SNe. Only at DEEP does accuracy dramatically drop to $\sim$ 40%, with confusion increasing to include both flare stars and XRB.
- **ESE**: ESEs are primarily confused with BGs and IDVs and classified with at least 90% accuracy in WIDE. MONTHLY with 14 observations maintains a classification accuracy of at least 65%.
- **IDV**: IDVs are classified with greater than 90% accuracy in all observing strategies except DEEP, and do not exhibit the characteristic decrease in accuracy over decreasing numbers of observations. Also, when inaccurately classified, its confusions are distributed among nearly all other classes, with a preference for SNe.
- **XRB**: XRBs exhibit the same classification signature of the flare stars, SNe and Novae. It is classified with $\sim$ 80% accuracy in WIDE and GP. Primary confusion is with RSCVn and SNe. Starting with LOG, accuracy decreases dramatically below 50%, with increasing confusion with Novae.

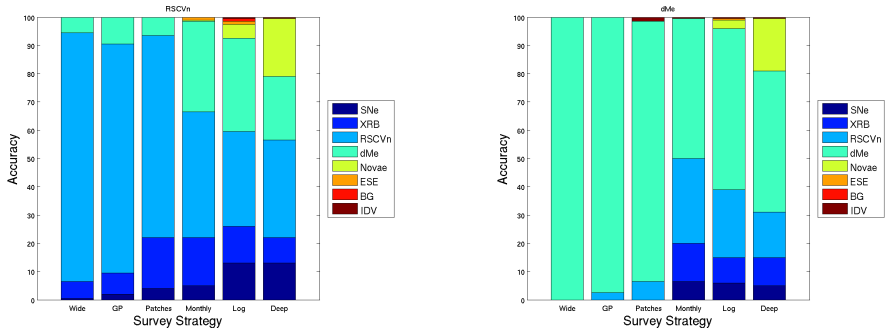
6.2 Online Classification

We compare classification results using 400-day-fixed, t-day-fixed, and t-day-adaptive classification (Figure 5(a)). Training on 400-day light curves while testing on t-day light curves is possible only using the stat or stat-cum feature sets which produce consistently-sized feature sets. The results in Figure 5(a) were generated using the stat-cum feature set.



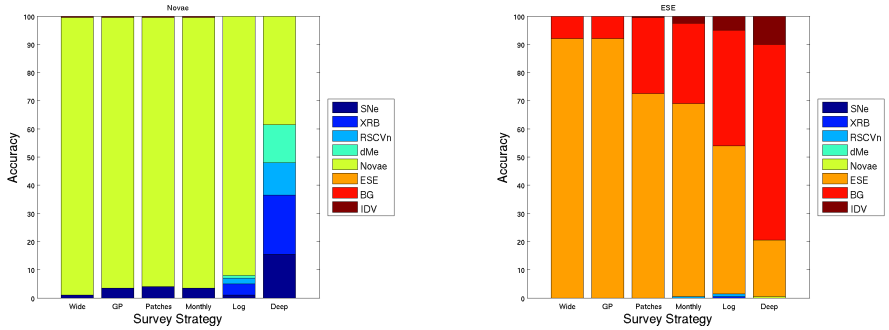
(a) BG

(b) SNe



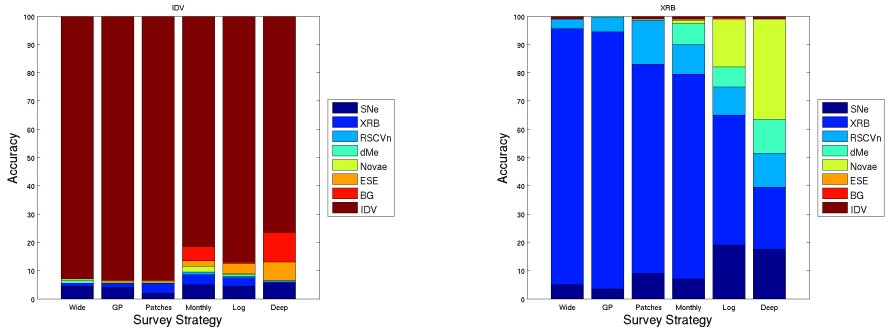
(c) RSCVn

(d) dMe



(e) Novae

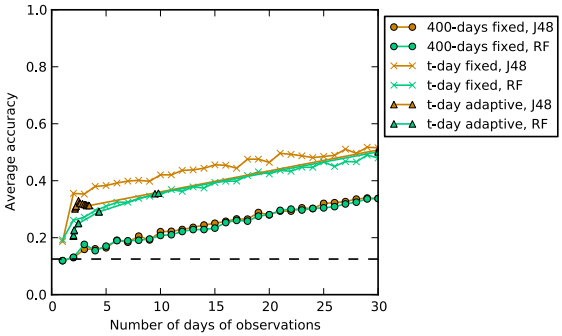
(f) ESE



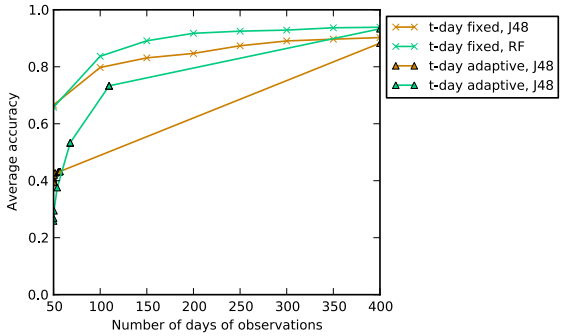
(g) IDV

(h) XRB

Figure 4: Class confusions using all-reps features and classifier SVM.



(a) Days 1 to 30



(b) Days 50 to 400

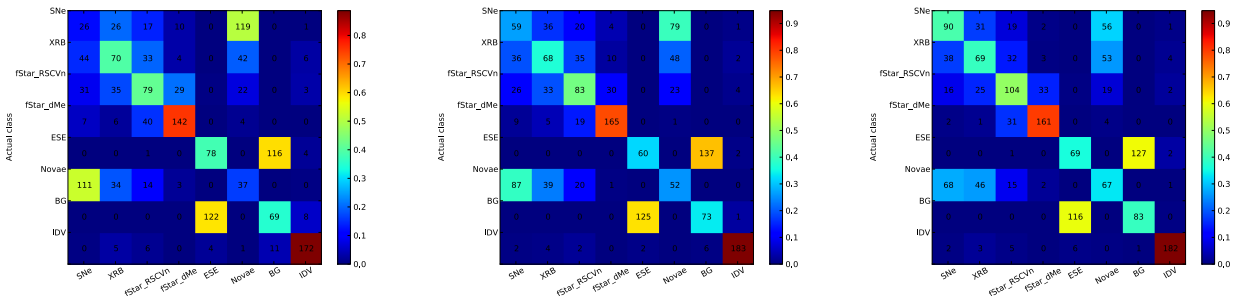
Figure 5: Online classification results using the stat-cum feature set, J48 and RF classifiers, and 10-fold cross validation. a) Compares the 400-day-fixed, t-day-fixed and t-day adaptive methods for days 1 to 30. The dotted line shows random classification accuracy. b) Shows the accuracy of the online classifiers correctly converge to the offline results at t=400 days.

When comparing 400-day-fixed with t-day-fixed, our key observations are:

- Overall, better performance was achieved when the model was trained with the t-day method. In this case, more lengthy observations did not improve test performance. This is perhaps not surprising since the features are more similar conceptually, spanning the same time period.
- When training on t-day light curves, average test performance reached $\sim$0.40. However, performance varied dramatically for different source types. IDVs and dMe flare stars achieved the best performance (0.70-0.80), while Novae, ESEs, and BGs had very low performance (0.00-0.30).
- On average, the biggest improvement in accuracy occurred during the first five days, after which accuracy increased more slowly for the next 25 days.
- J48 performed slightly better than Random Forest. However, the difference in performance narrowed with more days of observations.

We now analyze t-day-fixed against the t-day-adaptive method, where the t-day-adaptive method uses the first t-day classifier that classifies an example with sufficient confidence. In this setting, the number of observations required for a given source can vary, rather than committing to using the same number of observations (t) for all sources.

Figure 5(a) shows that the t-day-fixed classifier performed better than the t-day-adaptive classifier, somewhat to our surprise. Investigation of the results revealed that the posterior confidences produced by J48 and RF were often dramatically over-estimated. As a result, the adaptive classifier terminated too quickly, making a classification decision using only a few of the available 30 days of observations. This is why the t-day-adaptive results cluster strongly near the left side of the plot (low t values). As expected, the performance of the online classifiers converge to the offline results (Figure 5(b)).



(a) t=10 days (b) t=20 days (c) t=30 days

Figure 6: Online classification accuracy confusion matrices shown as heat maps (red = higher values, blue = lower values). J48, t-day training with stat-cum features. Overall accuracy at 30 days is 50%. The x-axis is the classified output class and the y-axis is the actual input class.

6.2.1 Class confusion

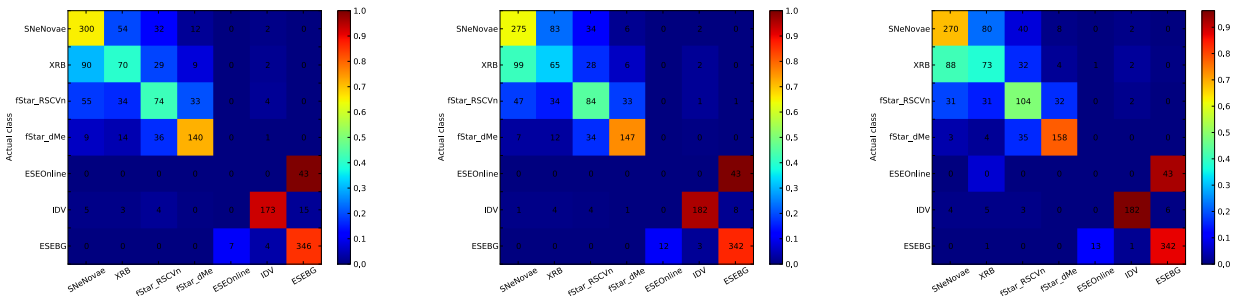
In this section, we explore the performance between different source types for the t-day-fixed training method with J48 classifier, which we have shown to be the best performing classifier. Figure 6 shows the confusion matrices at t=10, 20, and 30 days of observations. Our key findings are:

- IDVs show the best classification performance with accuracy of 92% at t=30 days.
- dMe flare stars show the second best classification performance with accuracy of 81% at t=30 days. Most of the misclassifications are attributed to RSCVn flare stars.
- BG sources are mainly confused with ESEs. This confusion did not decrease as number of observations increase from 10 to 30 days. There is almost no confusion between BG or ESE with any other classes even at t=10 days.
- SNes are most frequently confused with Novae. This confusion decreases with increasing number of days of observations. To a lesser extent, SNe are also confused with XRB and flare stars. However, this confusion does not decrease with more days of observations.

The majority of the incorrect classifications come from confusion between SNe and Novae, and between BG and ESE. This is not unexpected since the shape of the SNe lightcurve is similar to that of a Novae, and an ESE source without an event is indistinguishable from a BG source. Figure 7 shows the confusion matrix when we combine SNe and Novae into one class and ESE without significant event in the first 30 days and BG into one class. The ESEOnline class includes a subset of ESEs where flux changed by a substantial amount in the first 30 days. The overall accuracy at t=30 days of these seven classes is $\sim$ 70%, and improvement of $\sim$ 20% compared to the original eight classes.

7 Conclusions and Recommendations

We completed a study of VAST online and offline classification performance using simulated light curves from eight source types. Our offline classification experiments examined the impact of



(a) t=10 days (b) t=20 days (c) t=30 days

Figure 7: Online classification accuracy confusion matrices with SNe and Novae grouped as a class and BG and ESE (without significant events) grouped as a class. J48, t-day training with stat-cum features. Overall accuracy at 30 days is 70%. The x-axis is the classified output class and the y-axis is the actual input class.

different learning algorithms, light curve characterizations, and observing strategies over four levels of simulated source SNR. We summarize our offline conclusions as follows:

- Making use of all feature representations produces, on average, the best classification performance.
- All classification algorithms perform similarly when using best-performing feature representations.
- The concatenation of cumulative statistics (stat-cum) and two frequency characterizations (lsp and wlet), in conjunction with a support vector machine (SVM) classifier, achieved a best (on average) classification accuracy of $\sim 93\%$.
- Source types IDV and Novae classify with over 90% accuracy in all observing strategies except DEEP. Source types BG, SNe, dMe, Novae, ESE, IDV and XRB classify with over 90% accuracy in WIDE and GP.
- There is also natural confusion between source types BG and ESE, as well as among SNe, Novae, XRB, RSCVn and dMe.

Our online classification experiments tested performance using up to 30 days of observations with three learning approaches: train/test on archival light curves, train/test on t-day light curves, and adaptive-t classification. We summarize the conclusions from those experiments as follows:

- Training on t-day light curves performs better than training 400-day light curves.
- Both adaptive-t and (fixed) t-day methods performed better using a decision tree (J48) rather than a random forest (RF) classifier.
- The J48 classifier with cumulative statistical (stat-cum) features achieved the highest accuracy at all values of t. However, the accuracy reached on $t = 30$ days is only $\sim 50\%$ for the eight class classification problem.

- Source type IDVs and dMe flare stars achieved the best performance (0.70-0.80) at $t = 30$ days.

This memo demonstrates that archival classification of VAST transients and variables may be robust, but online classification will be challenging. We suggest the following paths forward to improve performance:

- Combine source types that are easily confused into groups, and adopt a hierarchical approach of discriminating among top-level groups, and then discriminating among source types within groups (using a classifier specialized for each group).
- Consult auxiliary sources where possible to augment classification performance.

The major challenge of estimating classification performance prior to the launch of ASKAP is the absence of real "VAST-like" light curves. Our simulations fell short of realistic scenarios in many ways, providing upper limits on true performance. In order to establish true expected performance, one can improve the simulations in the following ways:

- Employ a realistic distribution of source types (e.g., 10^5 BG for each SNe).
- For flare stars and ESEs, randomize the appearance of the transient event in relation to the background source on top of which it occurs.
- Simulate the effects of bad weather and instrument defects that result in missing observations.

Our future plans are to continue developing and evaluating our methods on both our simulated VAST data, as well as real radio and optical data sets, and compare our methods with those developed for similar real-time classification systems in optical astronomy (Djorgovski et al. [2011], Bloom et al. [2011]). Our goal is to improve both offline and online classification performance, and implement our best methods in the VAST pipeline.

8 Acknowledgements

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

References

Tara Murphy, Shami Chatterjee, and the VAST Collaboration. An ASKAP survey of variables and slow transients. *http://www.physics.usyd.edu.au/sifa/vast/index.php/Main/Documents*, 2009.

J. Banyer, T. Murphy, and the VAST Collaboration. VAST - a real-time pipeline for detecting radio transients and variables on the Australian SKA Pathfinder (ASKAP) telescope. *arXiv:1201.3130*, January 2012.

Joseph W Richards, Dan L Starr, Nathaniel R Butler, Joshua S Bloom, John M Brewer, Arien Crellin-Quick, Justin Higgins, Rachel Kennedy, and Maxime Rischard. On machine-learned classification of variable stars with sparse and noisy time-series data. *1101.1959*, January 2011.

J. D. Scargle. Studies in astronomical time series analysis. ii-statistical aspects of spectral analysis of unevenly spaced data. *Astrophysical Journal*, 263:835–853, 1982.

G. Wachman, R. Khardon, P. Protopapas, and C. R. Alcock. Kernels for periodic time series arising in astronomy. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part II*, pages 489–505, 2009.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995. ISSN 0885-6125.

J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. ISSN 0885-6125.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1), 2009.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

John C. Platt. *Fast training of support vector machines using sequential minimal optimization*, pages 185–208. MIT Press, Cambridge, MA, USA, 1999. ISBN 0-262-19416-3.

S.S. Keerthi, S.K. Shevade, C. Bhattacharyya, and K.R.K. Murthy. Improvements to platt's SMO algorithm for SVM classifier design. *Neural Computation*, 13(3):637–649, 2001.

Marie desJardins, James MacGlashan, and Kiri L. Wagstaff. Confidence-based feature acquisition to minimize training and test costs. In *Proceedings of the SIAM Conference on Data Mining*, pages 514–524, 2010.

S. G Djorgovski, C. Donalek, A. Mahabal, B. Moghaddam, M. Turmon, M. Graham, A. Drake, N. Sharma, and Y. Chen. Towards an automated classification of transient events in synoptic sky surveys. *arXiv:1110.4655*, October 2011.

J. S Bloom, J. W Richards, P. E Nugent, R. M Quimby, M. M Kasliwal, D. L Starr, D. Poznanski, E. O Ofek, S. B Cenko, N. R Butler, S. R Kulkarni, A. Gal-Yam, and N. Law. Automating discovery and classification of transients and variable stars in the synoptic survey era. *1106.5491*, June 2011.