Feature Based Astronomical Transient Classification

Final Honors Presentation

Peter Ashwell

Astronomy has been Transformed by Technology

Galileo (1564-1642)

ASKAP telescope array (2011)

The ASKAP telescope will produce 1 GB/s of data. The VAST project[1] aims to develop a pipeline that will analyse this data stream in real-time.

[1] http://www.physics.usyd.edu.au/sifa/vast/index.php/

Transients and Time Series

The goal of VAST is to develop a pipeline to detect and classify astronomical transient events in time series as early as possible.

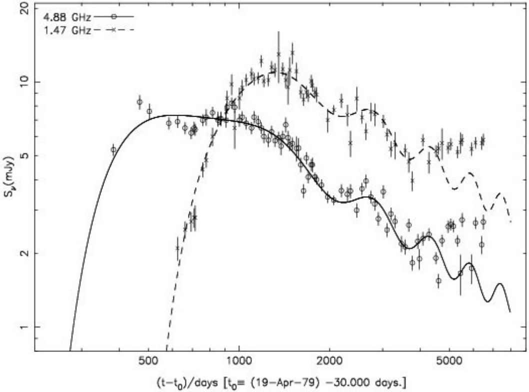


Figure: Real supernova time series data, in a couple of observing frequences. A model has been fitted over each frequency dataset.

Problem and Approach

My thesis addressed the transient classification component of the VAST pipeline. Key components of the work were

- Framing the problem as one of time series classification
- Performing an extensive literature review across application domains
- Developing an experimental framework with simulated transients and distortions typical to astronomical data
- Developing a feature based supervised classification scheme
- Using the framework to evaluate the classifier and to characterize the difficulties of transient classification

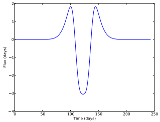
Research Questions

Through my approach I addressed the following research questions:

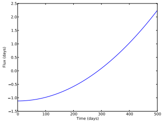
- Is supervised learning appropriate for astronomical transient classification?
- How do the various distortions, individually and combined, impact classification accuracy?
- What features cope best with what distortions?
- Is the classifier suitable for use in the VAST pipeline?

Simulated transients

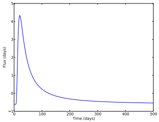
Existing data is either not of sufficient quantity for machine learning, not of the transient types of interest to VAST, or not of the data quality comparable to ASKAP. Simulated transient models were used, produced by Kitty Lo [3]



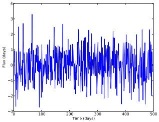
ESE



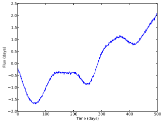
Nova



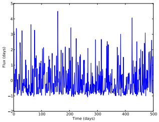
SNe



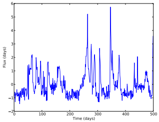
Noise



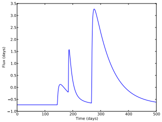
IDV



FSdMe



FSRSCVn



XRB

Simulated Distortions and the Dataset

Using these models I produced simulations of telescope data for a number of distortions at differing severities.

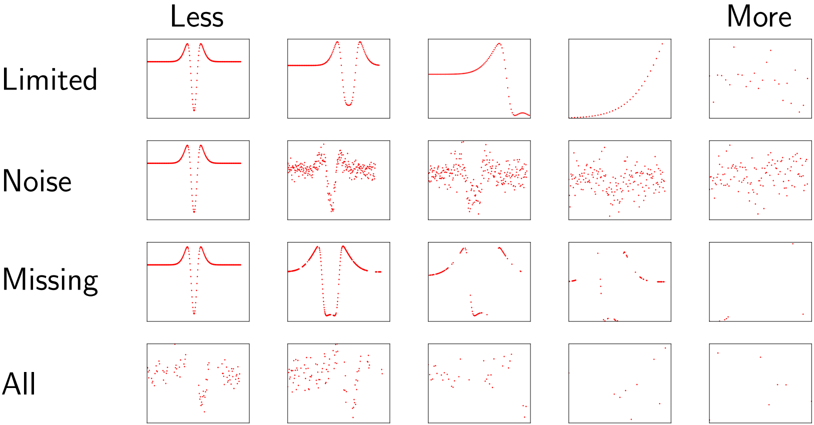


Figure: Distortions applied to ESE light curves

Background and Literature

Transient classification is a previously unexplored problem in machine learning or astronomy so a broad literature review of time series classification was performed:

Statistical models

- Gaussian processes [6]
- Hidden Markov Models

Distance measures

- Euclidean
- Dynamic time warping [8]
- Longest common subsequence

Features and transformations

- Linear segmentation [2]
- Piecewise aggregrate approximation
- Shapelets [9]
- Haar Wavelet Transforms [5]
- Lomb-Scargle periodograms [4]
- Statistical properties of flux distributions [7]

Classifier

Weka implementation of the Random Forest [1] supervised classifier with the following features:

- Haar wavelet coefficients
- Lomb-Scargle periodogram[2]
- Statistical properties of flux distribution
- Statistical properties of linear segmentation[3]
- Shapelets

For those features with no footnote I did the implementation.

[2] http://www.astropython.org/blog/2010/9/Question-period-finding-packages-in-python

[3] http://www.hackchina.com/en/cont/174005

Experimental Method

- Assumes a transient has been detected in a sliding window, passed to the classifier
- Dataset of 200 × 8 simulated transient classes
- Noise, missing data and a combination of both applied
- Cropping of time series to assess early classification
- Two experiments run for each distortion:
  - equally distorted training and test data
  - undistorted training and distorted test data
- 10-fold cross validation
- Results as F-Score and F-Score standard deviation across cross folds

Baseline results

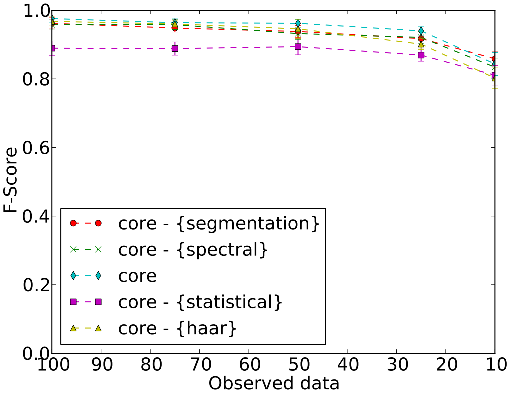
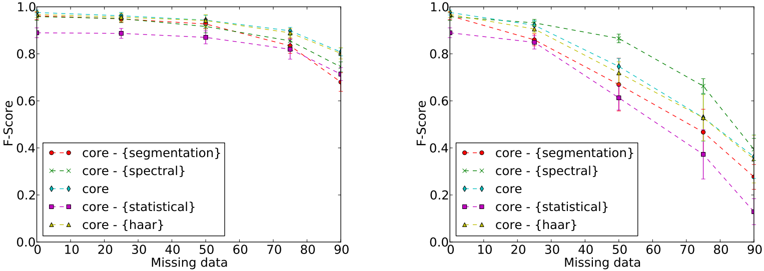


Figure: Classification performance as the percentage of each of the test cases observed is reduced. Classification is nearly perfect from 100% to 25% observed data with F-Scores above 0.95.

Missing data results

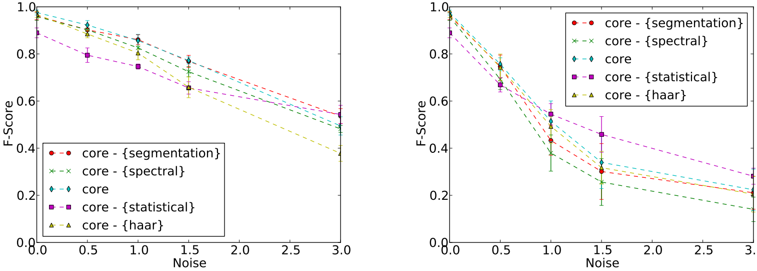


Distorted training data

Undistorted training data

Figure: F-Score under the missing data distortion. Classification stays above 0.9 F-Score until 75% missing data when using distorted training data. When using undistorted training data performance is significantly lower.

Noise results

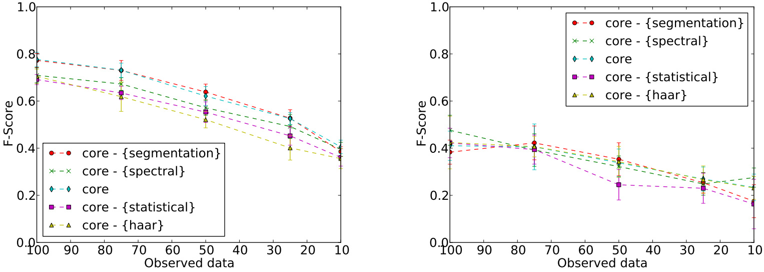


Distorted training data

Undistorted training data

Figure: F-Score under the noise distortion. When using distorted training data classification performance decreases steadily as noise increases. When using undistorted training data classification performance is much worse for the same amounts of noise.

Combined distortion results

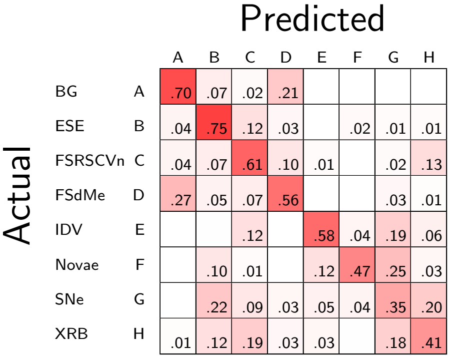


Undistorted training data

Distorted training data

Figure: Classification performance as the signal is cropped under both noise and missing data distortions. Classification performance when using data is below 0.8 F-Score for any amount of observed data. With undistorted training data it is below 0.4 F-Score.

Shapelet results



Predicted

Actual

		A	B	C	D	E	F	G	H
BG	A	.70	.07	.02	.21				
ESE	B	.04	.75	.12	.03		.02	.01	.01
FSRSCVn	C	.04	.07	.61	.10	.01		.02	.13
FSdMe	D	.27	.05	.07	.56			.03	.01
IDV	E			.12		.58	.04	.19	.06
Novae	F		.10	.01		.12	.47	.25	.03
SNe	G		.22	.09	.03	.05	.04	.35	.20
XRB	H	.01	.12	.19	.03	.03		.18	.41

Figure: Confusion matrices for the shapelet feature with an F-Score of 0.57. The shapelet feature can classify some classes (ESE, FSRSCVn) well.

The shapelet algorithm needs further work to give meaningful results for the missing data and noise distortions.

Analysis and Evaluation

- The Random Forest cannot classify accurately with the shift in feature values when using undistorted training data for any distortion.
- Signal clarity becomes an issue for noise after 1.0 and missing data at 90%.
- Compounded distortions give classification performance at 0.4 F-Score, not practical for the VAST pipeline
- The shapelet classification deserves further investigation as an additional feature
- Haar wavelets are good for dealing with noise. Statistical features are robust to missing data but are heavily affected by noise. Spectral features are affected by missing data.

Future work

- Investigating additional features to add to the classifier that will improve robustness to noise and missing data.
- Data proprocessing (smoothing, regression) to make the test data closer to the training data and reduce training/test feature value disparity.
- Modifying the shapelet algorithm to make it applicable to distorted data.

Conclusion

- Automated transient classification with the VAST pipeline has great scientific potential for astronomers, but is a difficult problem to solve.
- This research reveals two main difficulties in solving the problem in the context of the VAST pipeline:
  1. The lack of training data that is meaningful for incoming data with varying levels of distortions
  2. The compounded effects of noise and missing data on signal quality
- Future work in preprocessing the test data and adding additional features to improve classification under noise will improve classification performance and may make this approach suitable for VAST.

References

[1] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[2] X. Liu, Z. Lin, and H. Wang. Novel online methods for time series segmentation. *IEEE Transactions on Knowledge and Data Engineering*, pages 1616–1626, 2008. ISSN 1041-4347.

[3] Kitty Lo. Vast memo in prep. 2011.

[4] NR Lomb. Least-squares frequency analysis of unequally spaced data. *Astrophysics and space science*, 39(2):447–462, 1976. ISSN 0004-640X.

[5] R. Price, S. Vincent, and S. LeBohec. Haar wavelets as a tool for the statistical characterization of variability. *Astroparticle Physics*, 2011. ISSN 0927-6505.

[6] Carl E. Rasmussen and Christopher Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[7] J.W. Richards, D.L. Starr, N.R. Butler, J.S. Bloom, J.M. Brewer, A. Crellin-Quick, J. Higgins, R. Kennedy, and M. Rischard. On machine-learned classification of variable stars with sparse and noisy time-series data. *The Astrophysical Journal*, 733:10, 2011.

[8] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 26(1):43–49, 1978. ISSN 0096-3518.

[9] L. Ye and E. Keogh. Time series shapelets: a new primitive for data mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 947–956. ACM, 2009.

Shapelets

Provides a potentially useful classification feature whose value will change less than histogram or wavelet features as distortions are introduced.

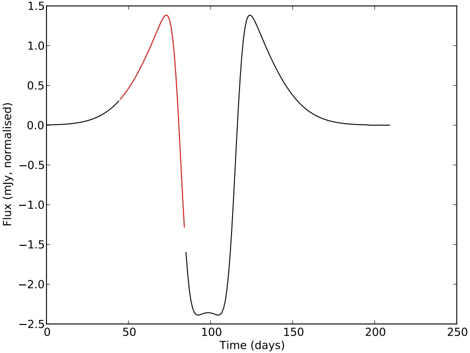
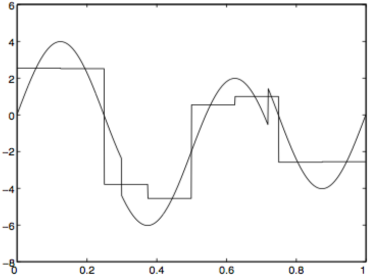
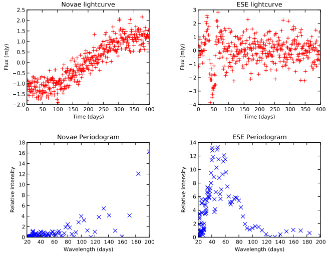


Figure: Shapelet for an ESE light curve from the dataset. The minimum Euclidean distance to a test light curve for the shapelet is used as a feature.

Wavelet transforms

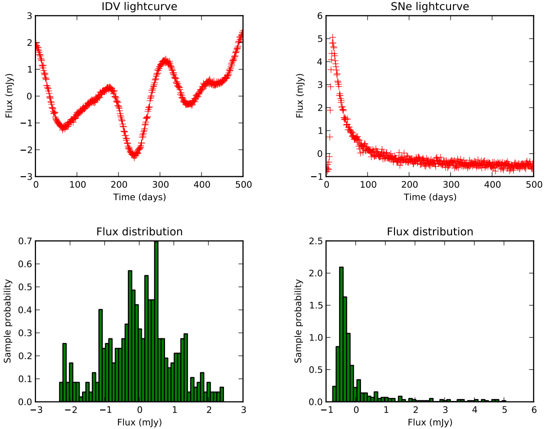




The coefficients of top 16 Haar wavelets and the top 5 strongest periodogram peak frequencies are used as features.

Statistical features

Applies various statistical measures including *Kurtosis* and *Skew* to histogram of z-normalised flux



Linear segmentation

The same statistical measures are applied to a histogram produced from a linear segmentation.

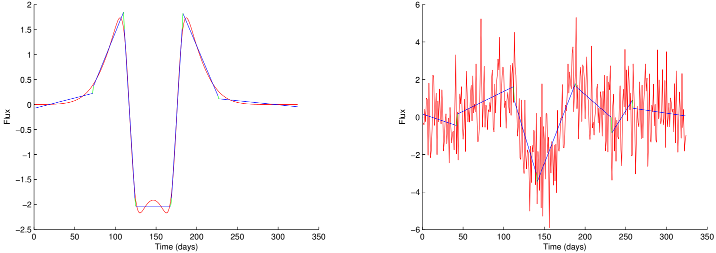


Figure: Linear segmentation of a time series from my dataset, and then for the same time series with 1.5 times its variance added as gaussian noise.

Classifier scalability

In the VAST pipeline classifier training is done offline. The running time of classifying with an already trained Random Forest is $O(\log n)$, where n is the number of features.

Lomb-Scargle periodogram	$O(L \log L)$
Haar wavelet transform	$O(L)$
Linear segmentation	$O(L)$
All statistical features	$O(L)$

Where L is the length of the time series in datapoints.