Identifying Contact Binaries in the Catalina Real-Time Transient Survey

Ricardo Elias Roche (Advisor: Dr. Peter Freeman, Department of Statistics)

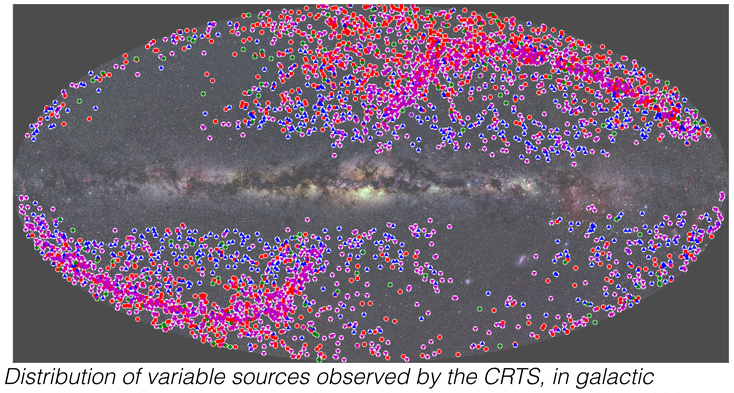
Introduction

The Large Synoptic Survey Telescope (LSST), situated in Chile, will observe the sky above the Southern Hemisphere for ten years, beginning in 2021. While most astronomical sources give off light steadily, some sources are 'variable'—their light output changes over time, either stochastically or periodically. Every spot in the southern sky will be observed by LSST roughly 1,000 times. This means that for every star or galaxy that LSST can observe, there will be a time series of estimated apparent luminosities. Vast numbers of new variable sources will be detected, far too many for other telescopes to observe and study. Thus astronomers must use the information present in the raw light curves to determine, in real time, which variable sources are worthy of further study. One solution would be to use an alert 'broker,' an algorithm that uses statistical and machine learning techniques to rank variable stars for follow-up observation. **An important aspect of such an alert broker algorithm is assigning stars to classes. In this poster, we compare the abilities of different classifiers to identify contact binary stars in the full sample of light curves.**

Data

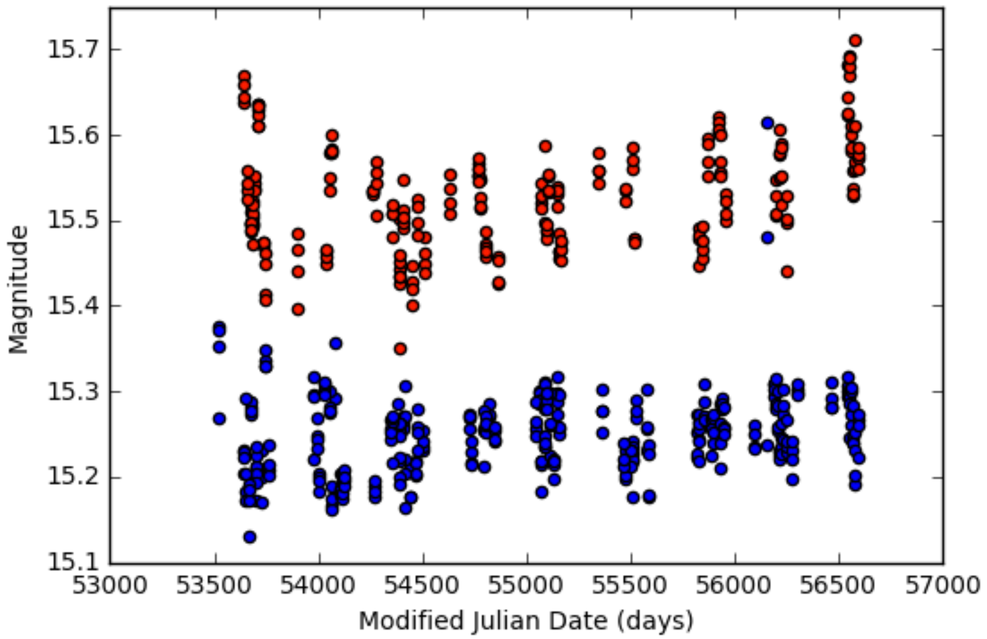
The Catalina Real-time Transient Survey uses three telescopes to cover 33,000 sq. degrees—just over 3/4— of the sky in order to discover variable and transient sources such as asteroids, supernovae, and Cepheid stars, etc. We analyze data from the Catalina Surveys Periodic Variable Star Catalog, which has light curves for 46,821 stars (Drake et al. 2014). These stars were observed a total of 13,712,786 times, for an average of 291 times each. The analysis by Drake et al. placed these stars into 17 different classes, the most populous being the contact binary class, with 30,593 members. For each observation, we use the following data.

Variable	Description
ID	unique identification number for each object
MJD	modified Julian date, i.e. the time of observation
Mag	the estimated V-band magnitude of the source at the observation time



Distribution of variable sources observed by the CRTS, in galactic coordinates. Some areas are without observation because dust within the Milky Way plane makes it difficult to observe distant sources on the disk, and the locations of the CRTS telescopes makes observing some areas easier than others. (http://nesssi.cacr.caltech.edu/catalina/transients.html)

Analysis

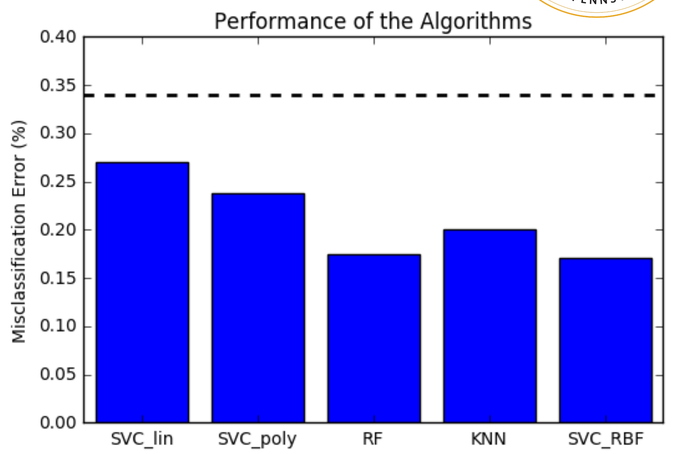


Light curves for variable sources: contact binary in red, other in blue. Magnitude is a logarithmically transformed measure of the brightness of a star. Because the stars are irregularly sampled, many conventional time-series analysis techniques are not applicable to these data.

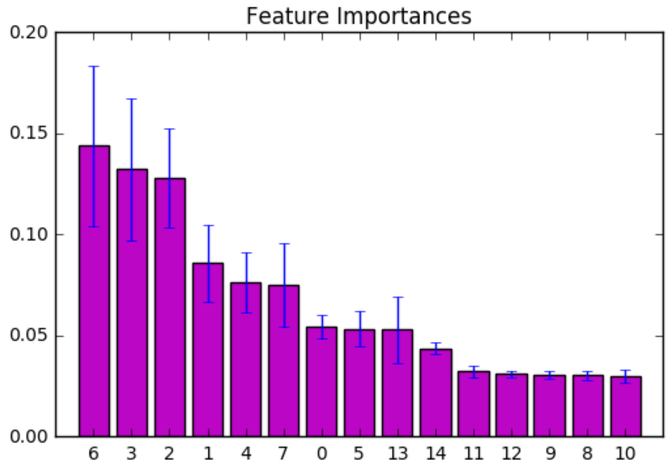
Light curves are high-dimensional objects, and thus it is conventional to reduce dimensionality by extracting features from each light curve. We extract the the following features defined by Richards et al. (2011):

#	Feature	Description
0	Mean	mean of the magnitudes
1	StDev	standard deviation of the magnitudes
2	BeyondStDev	percentage of the observations lying beyond one standard deviation from the mean
3	Skew	skew of the magnitudes
4	Kurtosis	kurtosis of the magnitudes, using Fisher's formula
5	Amplitude	difference between extreme magnitudes
6	Median Absolute Deviation	median discrepancy of the fluxes from the median flux
7	Median Buffer Range	percentage of fluxes within 20% of the amplitude from the median
8-12	Flux Percentile Ration Mid *n*	ratio of flux percentiles of the middle *n* percentiles over the 95th-5th percentiles (for these values of *n*: 20, 35, 50, 65, 80)
13	Percent Amplitude	largest absolute departure from the median flux, divided by the median flux
14	Difference Flux Percentile	ratio of flux of 95th-5th percentiles over the median flux

We use three different machine learning algorithms to classify the variable sources as either contact binaries or other class: k-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest (RF). Moreover, three different kernels of the SVM were used — linear, radial (RBF), and polynomial. We split the data into training and test sets to compare across performance across algorithms, and to determine the parameters that minimize misclassification error. The misclassification rates of the different classifiers on the test set are shown in the bar graph.



This bar graph shows the misclassification rate of each of the different classification algorithms. Note that a lower percentage implies better performance. The dashed line represents the maximum misclassification rate, 34%.



The Random Forest algorithm allows us to see the relative importance of each feature in predicting labels — that is, how much each feature contributes towards the classification. It is computed using permutation tests.

Conclusion

We find that SVM with a radial kernel yielded the best misclassification error, of 17.1%. Furthermore, we also find that the features of the data that best predict the label are the median absolute deviation, the skew of the magnitudes, the percentage beyond one standard deviation of the mean, and the standard deviation. While these features are useful in classifying contact binaries, it is worth noting that other features may be better for classifying different types of sources. Our work suggests that classifying transient objects based on their change in luminosity over time is a challenging analysis, since the 15 features extracted by the data reduce the misclassification rate by only 50%, which is too low to be optimally effective for the LSST.

References

- Drake, A. J., et al. 2014, The Astrophysical Journal Supplement Series, vol. 213, id 9
- James, G. et al. An Introduction to Statistical Learning: With Applications in R. Print
- Richards, J.W, et al. 2011, The Astrophysical Journal, vol 733, id 1.