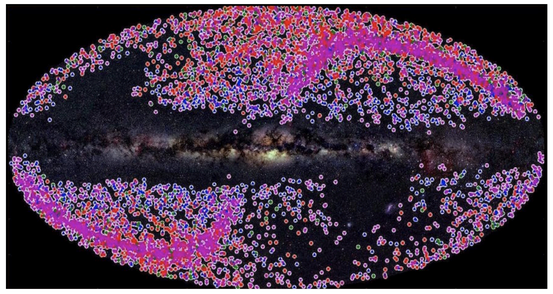
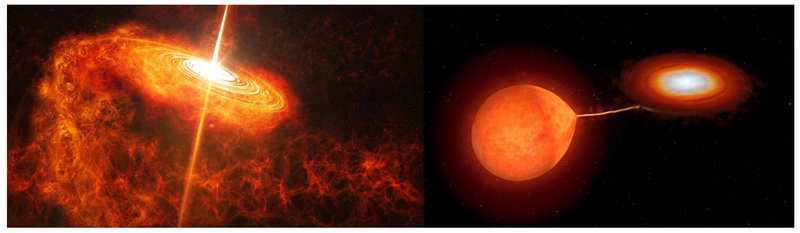


Classifying Blazars and Cataclysmic Variables from the Catalina Real-Time Transient Survey

Adrian Markelov, Kai Wen Wang, Yizhou Xu
{amarkelo, kaiwenw1, yizhoux} @ andrew.cmu.edu



Motivation

- There is a large and growing abundance of digital, synoptic astronomical surveys.
- Follow-up facilities are specialized to measure specific types of characteristics (distance, cadence, wavelengths, light intensity, etc.)
- Automated classification of detected phenomena is necessary to delegate follow-up facilities.

Introduction and Background

- Blazars are compact quasars associated with a supermassive black hole at the center of an active, giant elliptical galaxy.
- Cataclysmic Variables (CV) are binary systems consisting of two components: a white dwarf primary and a mass donor secondary.
- Although Blazars and CVs are very different in their physical appearance and behaviors, the observed light magnitudes are very similar in nature, and almost seemingly random.

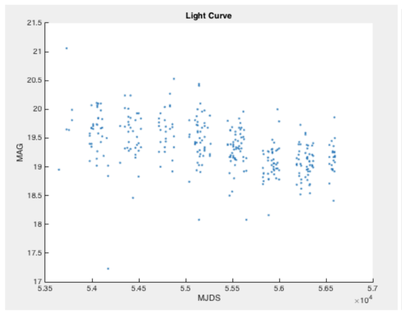


Figure 1: Example plot of raw Blazar data.

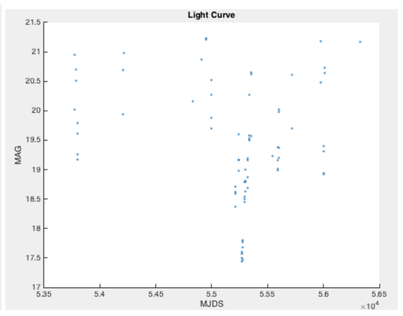


Figure 2: Example plot of raw CV data.

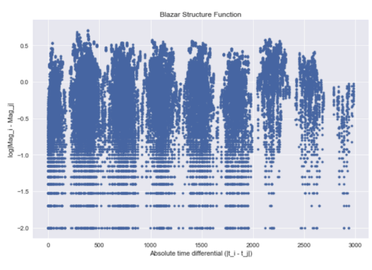


Figure 3: Plot of the structure function for Blazar 77.

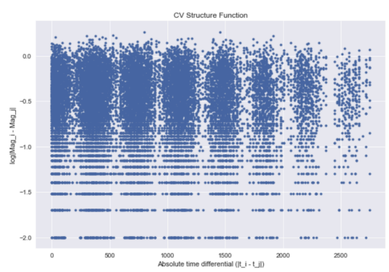


Figure 4: Plot of the structure function for CV 251.

Methods and Technical Details

Feature Transformations

1. Structure Function Transformation:

All transformations and models are performed on the basis of the structure function space.

Light Curve Function: $\phi : \mathbb{T} \to \mathbb{M} \quad s.t.$

$\mathbb{T} = \{\text{MJDT}\}$

$\mathbb{M} = \{\text{Light Magnitudes}\}$

Structure Function: $\psi : \mathbb{T}' \to \mathbb{M}' \quad s.t.$

$\mathbb{T}' = \{t' : t' = |t_i - t_j| \quad s.t. \quad t \in \mathbb{T}, i, j \in \mathbb{N}\}$

$\mathbb{M}' = \{m' : m' = \log_{10} |m_i - m_j| \quad s.t. \quad m \in \mathbb{M}, i, j \in \mathbb{N}\}$

2. Quantile Regression Transformation:

The quantile regression transformation estimates the distribution of the data along the log magnitude differential axis and reduces the dimensionality to 20.

$Q : [0, 1] \to \mathbb{R} \qquad Q(p) = \inf\{x \in \mathbb{R} : p \leq F(x)\}$

where F is the cumulative distribution function.

3. PCA Transformation:

a. Determine valid K dimensional space that preserves most of the information using explained variance

b. Use first K principal components as new feature space

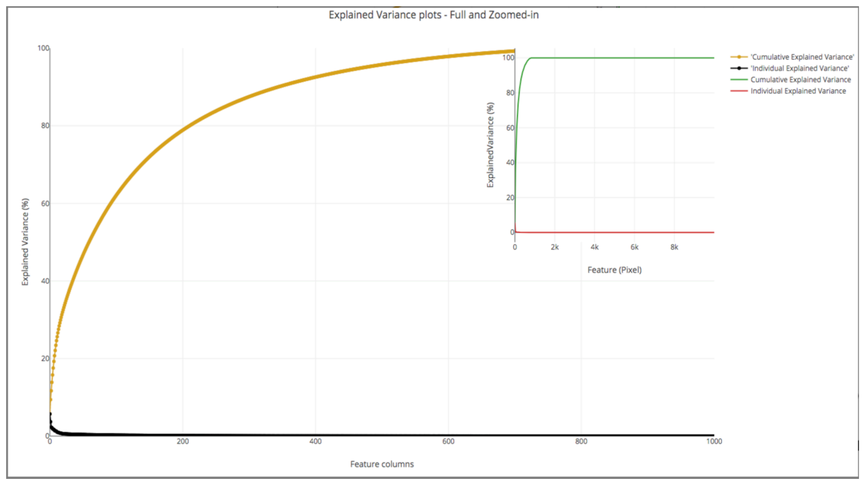
Classification Models

1. Basic models:

kNN, SVM, Adaboost, random forrests, neural networks

2. Deep Convolutional Neural Network:

Three convolution layers with kernel size 5^2 , stride 1 and 64,64,10 filters respectively, with ReLU activation and 0.5 dropout. Fully-connected layer for binary classification.



Experiments and Results

- Classification with basic models on features extracted from quantile regression (20-dim vector)

model	kNN	SVM	Adaboost	Random Forest	Neural net
accuracy	86.1	85.8	84.5	84.6	85.0

- Using a Balanced Bagging classifier, we achieved an accuracy of 86.8%. We show the confusion matrix:

	Predicted CVs	Predicted Blazars
Actual CVs	101	14
Actual Blazars	7	37

- Our model has an ROC AUC score of 86.0%.
- PCA results:
  - The explained variance function (bottom center) tells use that 92.6% of the variance/ information comes from the first 400 out of 10,000 principle components giving us a new 20x20 image.
  - The first eigenspace and the low dimensional image result from PCA are plotted here:

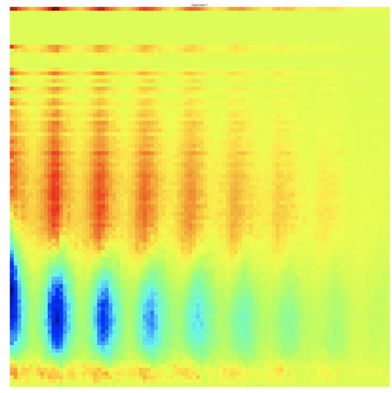


Figure 5: Plot of the first eigenspace of a structure function image.

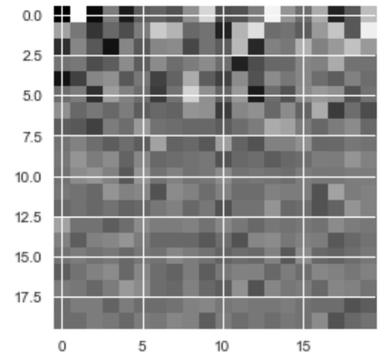


Figure 6: Plot of a low dimensional structure function transformed by PCA (100x100) -> (20,20).

- Using the PCA transformed images, our CNN model achieved around 90% accuracy.

Future Work and Discussions

- The extraction of good features is a major challenge. Overfitting is still an issue.
- Next steps:
  - Extract distributional features of each cluster of points (Figure 3,4) using Gaussian processes.
  - Signal processing techniques, Harmonic Analysis.

Carnegie Mellon University