ALLStars: Overcoming Multi-Survey Selection Bias using Crowd-Sourced Active Learning

Dan Starr[1], Joseph Richards[1,2], Henrik Brink[4], Adam Miller[1], Josh Bloom[1], Nathaniel Butler[3], J. Berian James[4], James Long[2], John Rice[2]
1 - Dept. Astronomy, U.C. Berkeley 2 - Dept. Statistics, U.C. Berkeley 3 - Arizona State University
4 - Dark Cosmology Centre, University of Copenhagen

Introduction

Developing a multi-survey time-series classifier presents several challenges. One problem is overcoming the sample selection bias which arises when the instruments or survey observing cadences differ between the training and testing datasets. In this case, the probabilistic distributions characterizing the sources in the training survey dataset differ from the source distributions in the other survey, resulting in poor results when a classifier is naively applied.

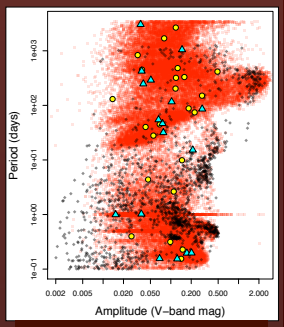
To resolve this, we have developed the ALLStars active learning web framework which allows us to bootstrap a classifier onto a new survey using a small set of optimally chosen sources which are then presented to users for manual classification.

Several iterations of this crowd-sourcing process results in a significantly improved classifier. Using this procedure, we have built a variable star light-curve classifier using OGLE, Hipparcos, and ASAS survey data and plan on bootstrapping onto SDSS and other active survey datasets.

Active Learning & Selection Bias

Active learning is a semi-supervised machine learning technique which incorporates a user or another resource capable of obtaining ground truth classifications for unlabeled data. The unlabeled items are algorithmically chosen to result in maximum improvement of the classifier when labels are found.

Active learning is well suited for adapting a classifier originally trained on one distribution of data (eg: a well understood historical survey) by both expanding the parameter space applicable to the classifier, and extending the classifier's prior probability distribution to be more representative of a new dataset.



Example of an active-learned sample distribution. Active learning samples are yellow and blue. Original Hipparcos & OGLE training dataset[3] are black and the ASAS testing dataset is red.

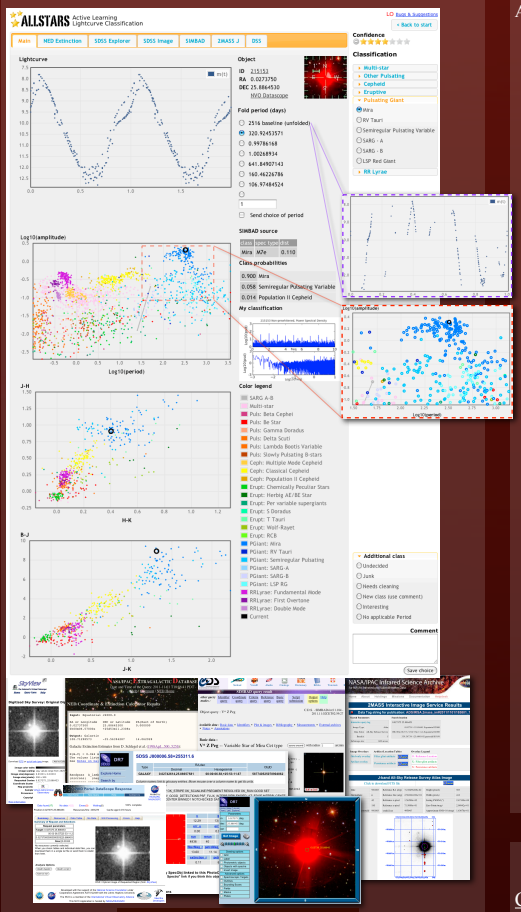
Class Discovery

During our active learning sessions it became apparent that several distinct red giant classes existed in the ASAS[4] dataset but were not represented in the OGLE & Hipparcos training dataset[3]. After defining new classes for OSARG-A, OSARG-B, and Long Secondary Period Pulsating Giants, the active learning process resulted in a classifier that could discern these types of science from the originally known & trained giant classes: Mira and Semiregular Pulsating Variables.

Active Learning algorithm logic

ALLStars: Active Learning Lightcurve Classification

A critical part of active learning is the ability to generate ground-truth labels for previously unknown sources. We've developed the ALLStars web framework to give users access to all available information about a source so they may make certain classifications. After reviewing the resources, the user can either make a science classification, flag the source as problematic, or skip to the next active learning source. They may also rate their confidence and store comments for that source.



Resources made available to a user of ALLStars.

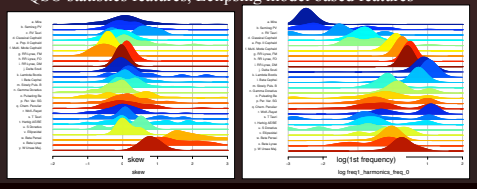
Once most users have classified the (~60) sources for a particular active learning iteration, final classifications are generated by combining user classifications using a proficiency weight. A user's classifications are then compared to the final classifications and their proficiency score is updated. This mechanism allows a user to improve their score as they become more proficient.

Random Forest & feature algorithms

After comparison of several machine learning frameworks, we found Random Forests to be the most effective for classifying our light-curve derived features.[2] Our Python framework wraps the R based randomForest classifier and has additional modifications to allow imputation of missing-value attributes.

A selection of the 84 attributes we use to describes a source:

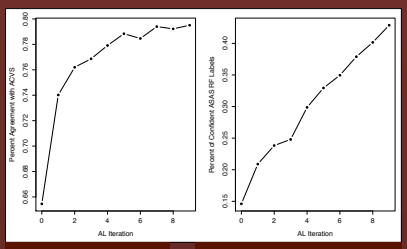
- 3 Lomb Scargle frequencies (pre-whitened, de-trended)
- model amplitude, relative phase, significance of each Lomb Scargle frequency and their 3 harmonics
- Amplitude, Flux percentile ratios
- model based slope percentiles
- Point2Point and median absolute deviation based features
- Color differences from the associated USNO NOMAD source
- statistics based features, count of n-day aliases
- QSO statistics features, Eclipsing model based features



Active Learning Results

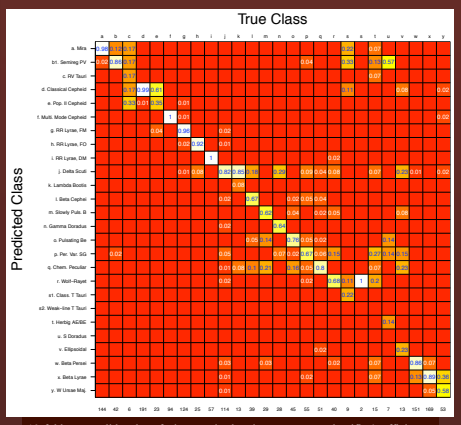
Our initial OGLE & Hipparcos training dataset was selected by Debosscher[3] and contains 1542 sources and 26 classes. The ASAS Catalog of Variable Stars[4] version 1.1 contains 50124 sources. 24% of these ACVS sources have "confident" machine learned classifications made by Pojmański.[4]

To determine the effectiveness of our active learning trained classifier, we compared our classifications against those in the ACVS catalog.



Our classifier's agreement with the ACVS labels, as a function of active learning iteration.

Percentage of ASAS data with confident (prob > 0.5) classifications, as a function of active learning iteration.



10-fold cross validated confusion matrix showing our current classifier's efficiency. This classifier was generated after 9 iterations of active learning with the ASAS dataset, and is applied to the original 1542 source OGLE & Hipparcos dataset. The cross validated error rate is 16.8%.

Berkeley CFTDI

Center For Time Domain Informatics http://cftd.info

- Collaboration of ~20 Professors, Post Docs, Grad Students
- UC Berkeley Astronomy, Statistics, Computer Science & Eng.
- Research:
  - Astronomy light-curve classification & real-time pipelines
  - Machine learning applied to GRB, SN, QSO problems
  - Statistics: classifying attributes with errors; outlier detection
- Upcoming conference on Online/Streaming Machine Learning:
  - May 7-11, 2012. U.C. Berkeley, CA, USA
  - http://cftd.info/home/2012-conference
- Contact us if interested in speaking at CFTDI's Seminar Series

DotAstro: Time Domain Astro. Warehouse http://dotastro.org

- Public website containing ~50000 downloadable light-curves
- From ~90 papers, has ~150 science classes

Contact: Dan Starr
dstarr@astro.berkeley.edu

References

1. Richards et al., (2011) *Active Learning to Overcome Sample Selection Bias: Application to Photometric Variable Star Classification*. arXiv:1106.2832, ApJ 743, 1
2. Richards et al., (2011) *On Machine-learned Classification of Variable Stars with Sparse and Noisy Time-series Data*. arXiv: 1101.1959, ApJ 733, 10
3. Debosscher et al., (2007) *Automated supervised classification of variable stars. I. Methodology*. arXiv: 0711.0703, A&A 475, 3, 1159
4. Pojmański, G., (1997) *The All Sky Automated Survey*. Acta Astron., 47, 467
5. Settles, B. (2010) *Active Learning Literature Survey*. CS Tech. Rep. 1648, University of Wisconsin-Madison