From Raw Data to Clean Catalogs

Automatic classification of variable stars in the MACHO survey

Roni Khardon
Tufts University
roni@cs.tufts.edu

Gabriel Wachman, Tufts University
Pavlos Protopapas, Harvard CfA
Charles Alcock, Harvard CfA

The Project & Talk Outline

- OGLE catalog as clean training data
- Completely automated system for processing, filtering, and classifying stars in MACHO survey
- Some interesting challenges and design decisions
- Lessons, open questions and future steps (computer science perspective)

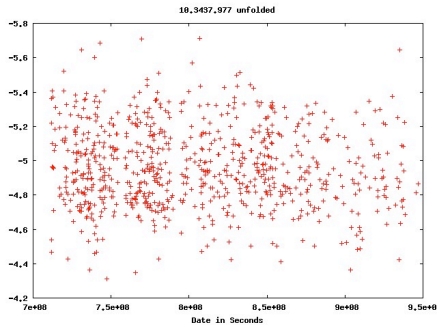
The Problem

- MACHO survey images processed to produce time series data B and R bands
- Data for 25,309,792 stars
- We are interested in variable periodic stars especially of types:
- Cepheid, RR Lyrae, and Eclipsing Binary
- A tiny fraction of the stars in survey

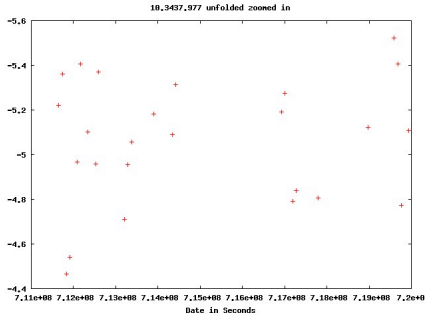
The Problem

- MACHO survey images processed to produce time series data B and R bands
- Data for 25,309,792 stars
- We are interested in variable periodic stars especially of types:
- Cepheid, RR Lyrae, and Eclipsing Binary
- A tiny fraction of the stars in survey
- We have OGLE data for similar fraction

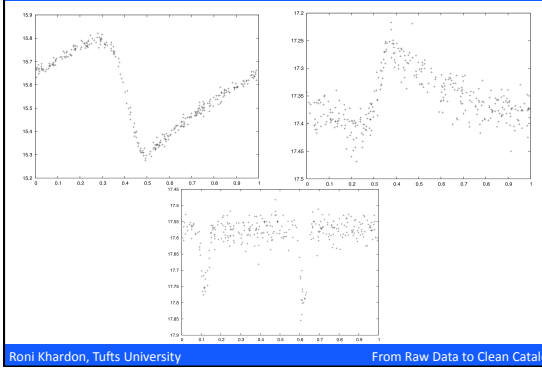
Raw Data



Raw Data (zoom in)



"Clean" Folded OGLE Data



How we Classify: SVM

- 3 numerical features:
  Period, Magnitude, Color
- Kernel for time series

$$K_1(x,y) = \sum_s e^{\gamma < x, y_{+s} >}$$

- Combined Kernel (similarity function)

$$K(x,y) = \sum_s e^{\gamma < x, y_{+s} >} + (p_x p_y + m_x m_y + c_x c_y)$$

How Good Can it Get?

Cross validation on OGLE (using OGLE periods)

	Ceph	EB	RRL
Ceph	3095	16	7
EB	22	2069	3
RRL	4	3	7036

Cross validation on OGLE (using our periods train/test)

	Ceph	EB	RRL
Ceph	3078	64	13
EB	30	2018	6
RRL	13	6	7027

Processing pipeline for MACHO

- Start with ~25M stars
- Filter: non-variable, too sparse
  ➔ ~8M stars left
- Filter: non-periodic (+ find period)
  ➔ ~62K stars left

These are stars of interest likely variable and (somewhat) periodic

Processing pipeline for MACHO

- Remove stars "too far" to Inspect group
  ➔ ~38K stars left
- [Train SVM classifier on OGLE data]
- Classify remaining stars into:
  Ceph, EB, RRL, or "Abstain"

Outcome for MACHO

- Ceph: ~7.5K stars
- EB: ~11K stars
- RRL: ~17K stars
- Abstain: ~2K stars

- Inspect: ~32K stars
- Discover several 1000s new stars in each class over previous work

Outcome for MACHO

- Ceph: ~7.5K stars
- EB: ~11K stars
- RRL: ~17K stars
- Abstain: ~2K stars

- Inspect: ~32K stars
- Discover several 1000s new stars in each class over previous work

Inspect and Abstain

- Both groups useful for further study
- Inspect group:
  - Likely to be interesting variable stars that need further examination.
  - Not likely to be in our 3 classes
- Abstain group:
  - Likely to be in our 3 classes
  - Learned model not sure about prediction

Discussion points/challenges

- Focus on end-to-end system
- Plan to publish data as catalog and code as tool for others
- Reliable large scale computing platform can be an issue
- Domain specific parameters are crucial (Kernel and features)

Discussion points/challenges

- Period finding is a major challenge
- Should we model the expected behavior (using generative models)?

- Early detection of events

- Statistical models for Inspect and Abstain
- Probabilistic classifiers? or SVM?

From Raw Data to Clean Catalogs

Automatic classification of variable stars in the MACHO survey

Roni Khardon
Tufts University
roni@cs.tufts.edu

Gabriel Wachman, Tufts University
Pavlos Protopapas, Harvard CfA
Charles Alcock, Harvard CfA