# Classification

Laurent Eyer
Observatoire de Genève
Wednesday 6th of July 2005

# Plan of the talk

- Previous work with large database
- Recent automated studies
  - Bayesian classifier
  - SOMs
  - Support Vector Machine

# Variable Star Automatic Classification

## Status of previous works

- Hipparcos (Geneva-Cambridge)
  - Light curve analysis
  - No real systematic classification
- OGLE, MACHO, EROS
  - Extraction of specific objects (RR Lyrae, Cepheids, Eclipsing binaries, etc...), but no global classification
- ASAS (All-Sky-Automated-Survey, G. Pojmanski)
  - Projection on selected 2D plane (selection was manual, semi-automated)

No automated classification

# Recent automated studies

Gaia needs full automation

- Bayesian Classifier: Eyer & Blake (2002, 2005)
- Neural Network: Belokurov et al. (microlensing 2003, transients 2004), Brett et al. (2004), Finney et al. (Novae identification, 2005)
- Self-Organising maps: Belokurov et al., Naud & Eyer
- Support Vector Machine: Willemsen & Eyer (2005)

# Bayesian Classifier
# Eyer & Blake

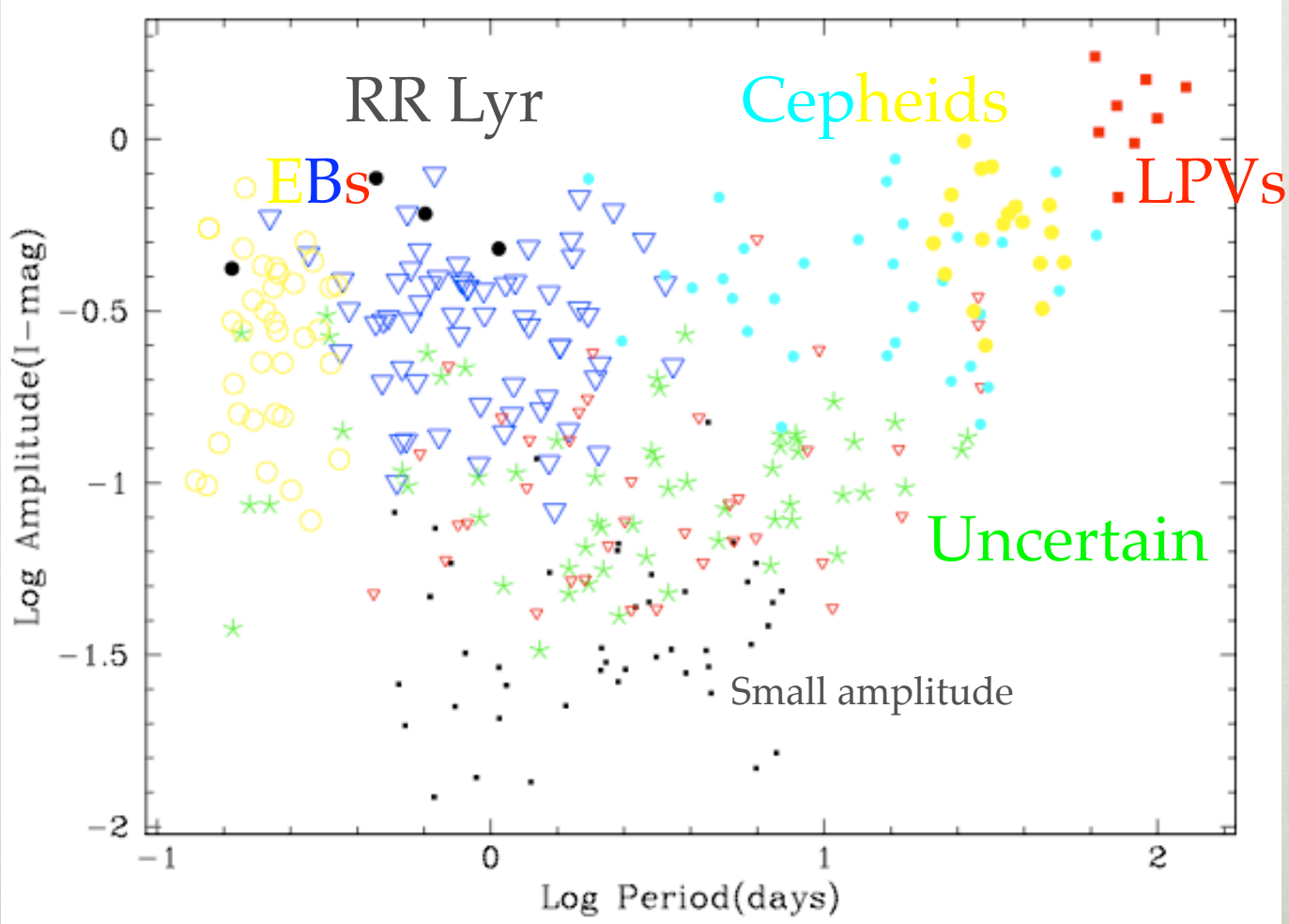
- All Sky Automated Survey (ASAS)
- Modest number of objects: 1700 stars
- One of the First real global automated classification!
- Error level 7%

# Example of the classification

1)Per
2)Amp

Fourier:

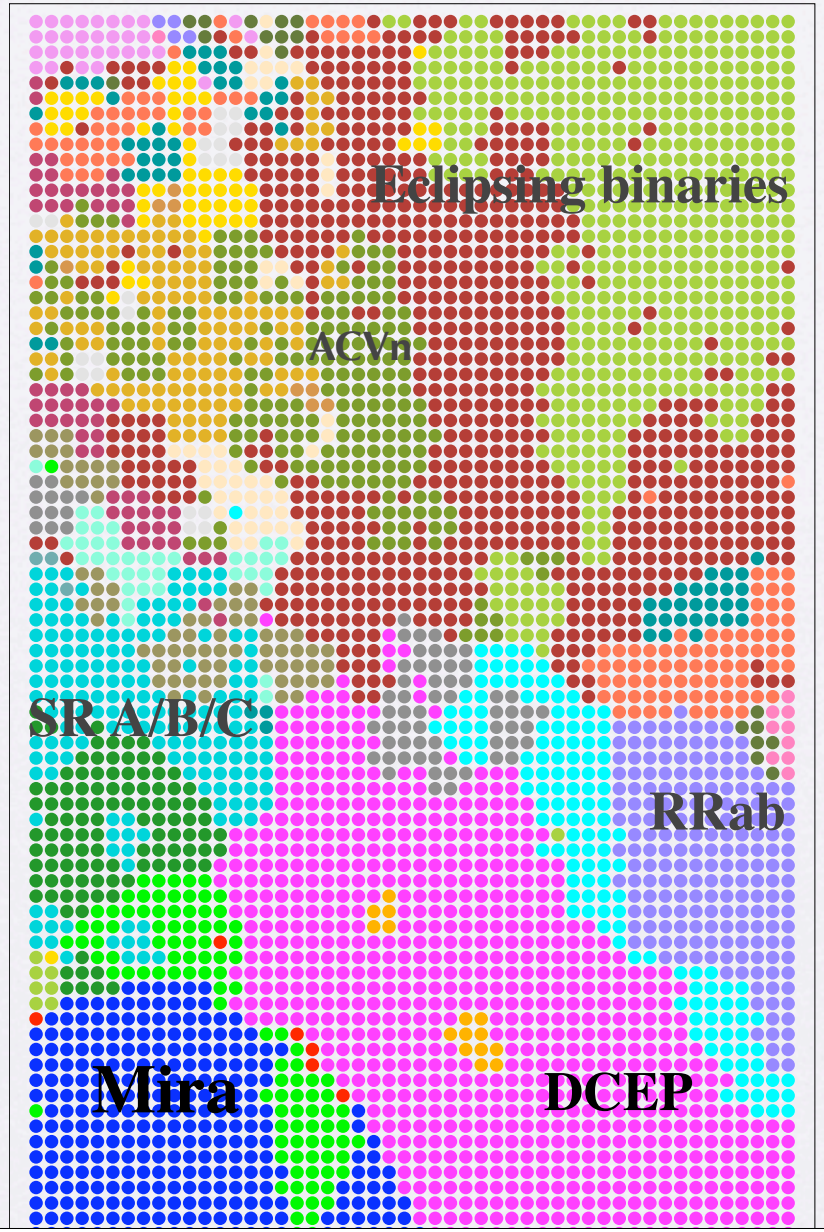
3)phi21
4)R21



# Self Organising maps Belokurov

- Go to Vasily Belokurov's web-site

# Self Organising maps




ACYG
ACV
BCEP
BY
CWA
CWB
DCEP
DCEPS
DSCT
EB
EA
ELL
EW
E
GCAS
M
RRC
RRAB
RS
SRB
SR
SARV
SRC
SRA
SPB
SXARI
SXPHE

## On Hipparcos data

A. Naud and L. Eyer

Features:

- Period
- Amplitude
- Colour V-I
- Skewness

# Support Vector Machines Willemsen & Eyer

- Why Hipparcos data?
  - Selection of Training set
  - Sampling peculiar (Number of measurements variable, sampling different from one star to an other)

Hipparcos:

4486 stars with variable types

Nomenclature uniformisation

Classes rejected and retained:

| Variability class | | |
|---|---|---|
| rejected | retained | # of stars per class |
| BE | ACV | 170 |
| BY | ACYG | 48 |
| BY+UV | BCEP | 59 |
| CEP | DCEP | 188 |
| CW | DSCT | 111 |
| CWA | EA | 472 |
| CWB | EB | 324 |
| DCE | ELL | 47 |
| DCEPS | EW | 113 |
| E+ZAN | GCAS | 198 |
| EA+BC | I | 517 |
| EA+DSC | L | 356 |
| ELL+XF | M | 190 |
| FKCOM | RRAB | 75 |
| NC | RS | 68 |
| NL | SPB | 91 |
| NL+ZZ | SRA | 42 |
| NR | SRB | 148 |
| PVTEL | | |
| RCB | | |
| RV | | |
| RVA | | |
| RVB | | |
| S | | |
| SARV | | |
| SDOR | | |
| SPB | | |
| SR+ZA | | |
| SR:/PN | | |
| SRA+E | | |
| SRC | | |
| SRD | | |
| SXARI | | |
| SXPHE | | |
| UV | | |
| WR | | |
| XNG | | |
| ZAND | | |

18 classes

# Selection of features

51 features:

- B-V, V-I
- skewness
- 10-percentiles median subtracted (d1-d9)
- 40 bins Fourier envelope

Principal Component analysis to reduce the dimensionality of the problem

# Confusion table

| | true | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **predicted** | ACV | ACYG | BCEP | DCEP | DSCT | EA | EB | ELL | EW | GCAS | I | L | M | RRAB | RS | SPB | SRA | SRB |
| ACV | 42 | 2 | 6 | | 1 | | 6 | 9 | 1 | 15 | 2 | 1 | | | | 36 | | |
| ACYG | | 4 | | | | | 2 | | | 1 | 2 | 2 | | | | | | |
| BCEP | 2 | | 3 | | | | | | | 1 | 1 | | | | | | | |
| DCEP | | | | 65 | | 1 | 1 | 1 | | | 2 | 2 | | | | | 1 | |
| DSCT | 1 | | 2 | | 35 | 1 | 5 | 1 | | 1 | 1 | | | 2 | | 1 | | |
| EA | 1 | | | | | 107 | 20 | | | 2 | 2 | | | | | | | |
| EB | 2 | | | 1 | 1 | 12 | 60 | 2 | 11 | 11 | 4 | 2 | | | | | | |
| ELL | | | | | | | | | | | | | | | | | | |
| EW | | | | | 1 | | 5 | | 28 | | | | | | | | | |
| GCAS | 5 | | 1 | 1 | | | 2 | | | 32 | 3 | 2 | | | | | 1 | |
| I | | | | 3 | 1 | 4 | 4 | 3 | | 5 | 124 | 71 | 1 | | 1 | | 1 | 11 |
| L | | | | 2 | | | | | | | 16 | 34 | | | | | 6 | 16 |
| M | | | | 1 | | 1 | | | | | 1 | 1 | 63 | | | | 5 | |
| RRAB | | | | | | | | | | | | | | 26 | | | | |
| RS | | 1 | | | | 2 | 1 | 1 | 1 | | 6 | 5 | | | 20 | | | |
| SPB | | | 1 | | | | 2 | 1 | | 1 | | | | | | 1 | | |
| SRA | | | | | | | | | | | 1 | 2 | | | | | 1 | 2 |
| SRB | | | | | | | | | | | 4 | 10 | | | | | 3 | 23 |
| **TP [%]** | **79.2** | **57.1** | **23.1** | **89.0** | **89.7** | **83.6** | **55.6** | **0** | **68.3** | **46.4** | **73.4** | **25.8** | **98.4** | **92.9** | **95.2** | **2.6** | **5.6** | **44.2** |
| objects/class | 53 | 7 | 13 | 73 | 39 | 128 | 108 | 18 | 41 | 69 | 169 | 132 | 64 | 28 | 21 | 38 | 18 | 52 |

Table 2: Confusion matrix for the classification results based on the original dataset and without class weighting. Shown are the numbers of objects in the true versus the predicted classes. For better clarity, only non-zero entries are shown. The line **TP** shows the number of true positives for each class, i.e. the percentage of correctly identified stars. The last line shows the number of objects per class in the validation set. Summing up the numbers on the diagonal of the confusion matrix and dividing by the total number of objects in the validation set (1071) yields an estimate of the overall classification performance. In this case, we find 62.4 % of correctly identified objects.

# Conclusion on SVM

- Classification performance 60%-80-98%
- Training set not good enough, class ill defined
- No substantial improvements in dimensionality reduction (with PCA)
- Prime importance: Confusion tables, estimations of false negatives and false positives

# ACTIONS

- Continue work on Hipparcos with SVM: define a better training set
- Include classification for Grid
- Benchmark for classification methods
  - completeness
  - flase positives, negatives