Data mining theory applied to the classification of light curves based on the HIPPARCOS catalogue

Luis M. Sarro, Frédéric Vincent, Amelia Bayo
Dpt. Inteligencia Artificial (UNED)

1.- Principles of data mining

- ✗ Extract knowledge from huge data warehouses (attributes+instances) of high dimensionality. Thousands of new classified objects as a by-product of the EXO field.
- ✗ Involves preprocessing, mining and validation/evaluation
- ✗ Mining encompasses tasks such as classification, regression, time series prediction, clustering, data completion, correlation detection, association rules discovery, outlier detection...
- ✗ Based mostly on statistics combined with information theory and...

In the following, we concentrate on a prototype developed in the last few months in order to classify COROT light curves based on the published catalogue of the HIPPARCOS mission.

2.- Data description

- We have tried to construct a pattern recognition agent using only the information in the HIPPARCOS catalogue that will also be available (somehow) in the COROT data: photometric time series and V-I colour.
- The first preprocessing stage consisted in deriving periods, amplitudes and fourier coefficients and ratios as well as a light curve folded in phase, completed (via a Self-Organized Map) and rebinned (in 50 phase bins) as described in the 7th COROT week in Granada.
- Time series analysis was performed following Vanicek's least squares method (1971) + Scargle's statistical significance tests (1981,1982). Periods were further refined using Phase Dispersion Minimization or the RGO Hipparcos method.

2.- Basic data statistics

- A total of 2712 objects in HIPPARCOS variability Annex 1
- A total of 526 objects in the test set belong to VA1 and cannot be used for training. **[key concepts: overtraining/generalization]**
- A total of 352 unclassified objects in VA1
- 26 broad classes of periodic variability with a significant fraction totally or almost depopulated (e.g. SX Arietis, PV Telescopii, FK Comae Berenices; see sect 2.04 of the catalogue).
- Initial number of attributes:
  4 (period, amplitude, V-I colour, mono/multiperiodicity)
  +50 light curve bins
  +16 fourier coefficients and ratios

2.- Data description

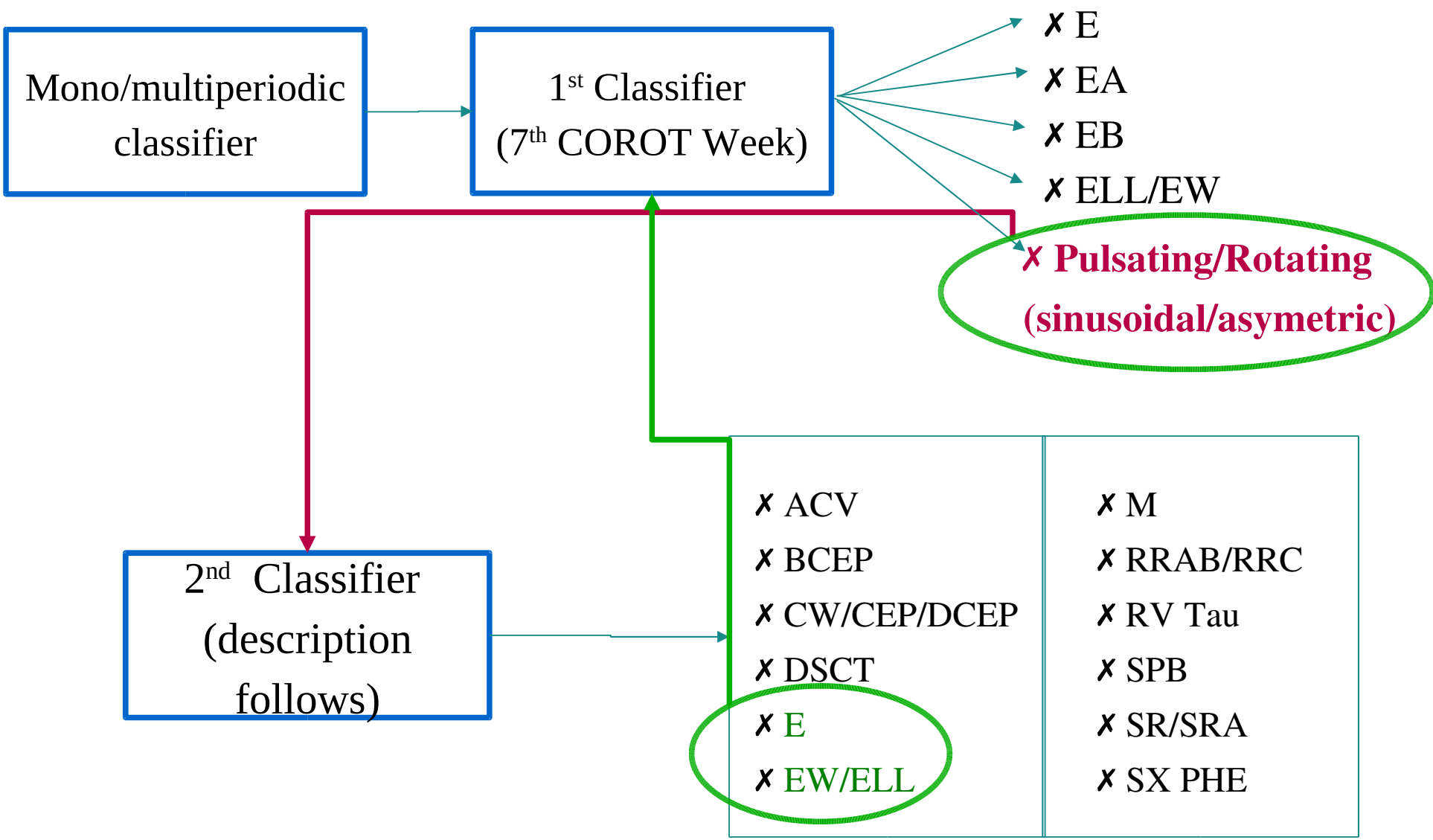
The second stage consisted in applying several techniques in order to determine the optimal attribute set for classification by considering:

- ✗ Wrapping
- ✗ Information Gain
- ✗ Principal Component Analysis

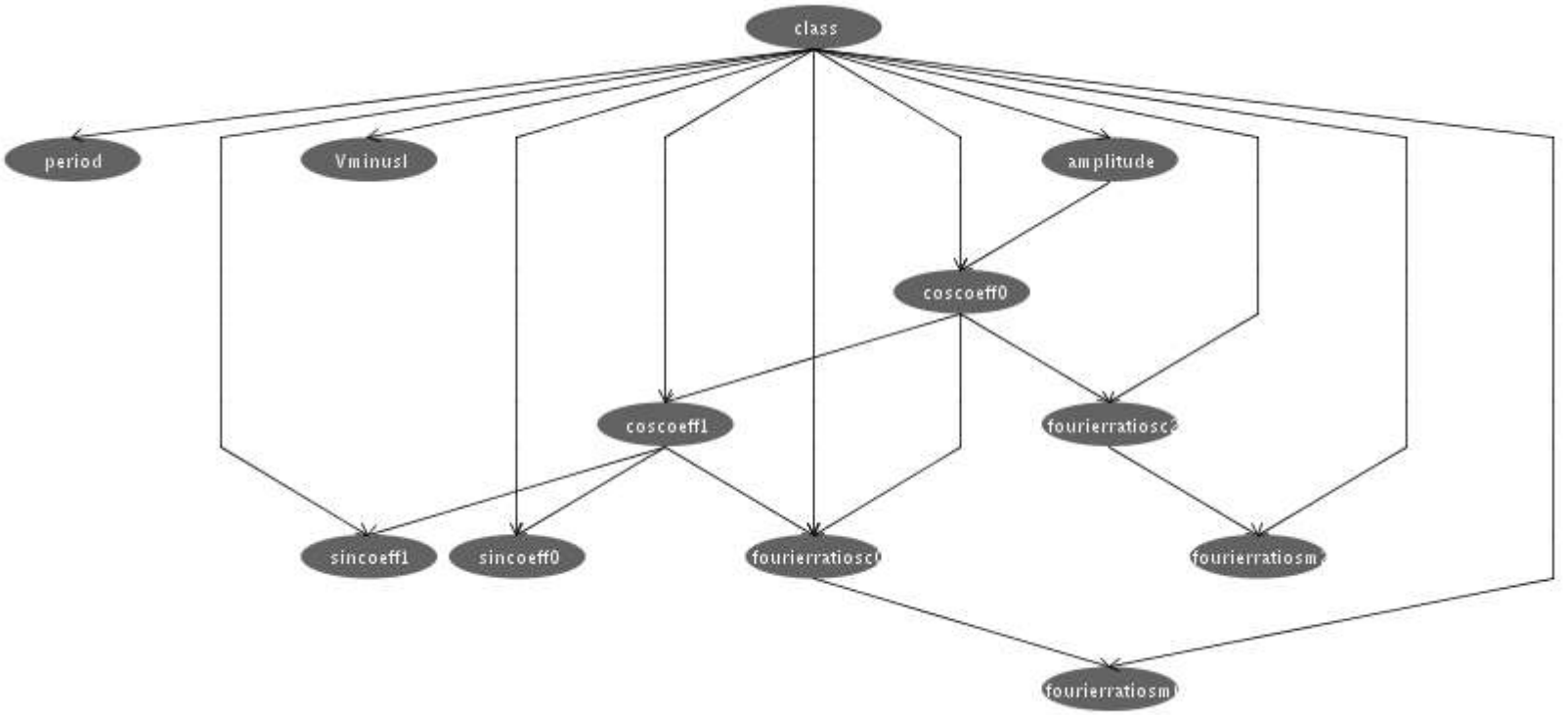
combined with different search methods (rankers, greedy best-first and genetic search)

The resulting ordered attribute set used for training was: period, amplitude, V-I , 1st and 2nd fourier sine coefficientes, 1st and 2nd fourier cosine coefficients and several coefficient ratios.

2.- Knowledge flow



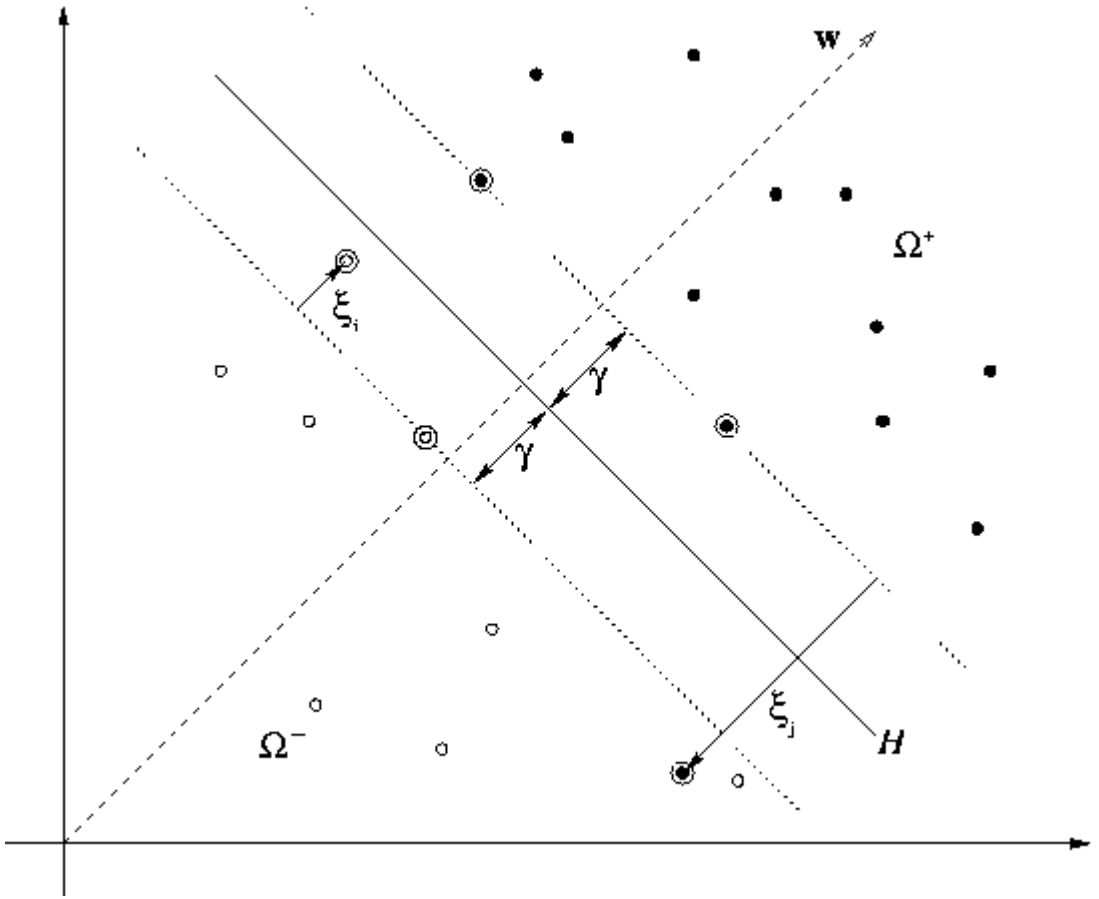
3.- Classification approaches I: Bayes Net



Bayes net constructed using the K2 search algorithm and metrics (although alternative metrics were tried such as MDL, AIC and entropy alone). Conditional probability distribution was computed with a maximum likelihood estimator.

Alternative search methods (Simulated Annealing and genetic search) were also explored.

3.- Classification approaches II: Support Vector Machines



Support Vector Machines try to find the optimal separating hyperplane following the Structural Risk Minimization principle.

We have used polynomial (degrees 1 and 2) and radial basis functions kernel mappings onto higher dimensional feature spaces.

3.- Classification approaches III: Bayesian Neural Networks

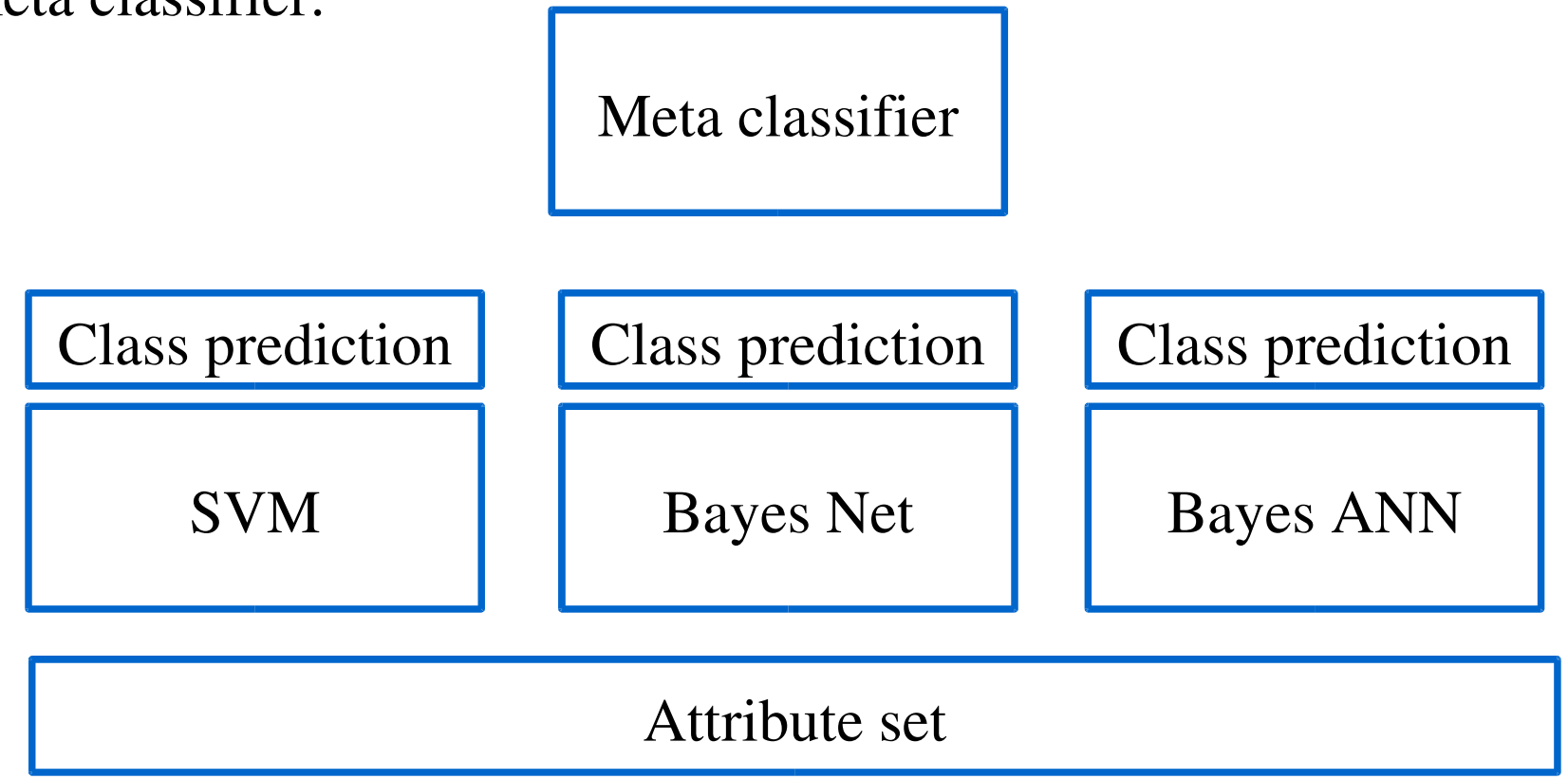
- We use Markov Chain Monte Carlo (MCMC) methods to generate an ensemble of multilayer perceptrons whose outputs are combined according to

$$P(C_{n+1}|\boldsymbol{x}_{n+1}, \mathcal{S}_{\text{train}}) = \int P(C_{n+1}|\boldsymbol{x}_{n+1}, \boldsymbol{\theta}) \cdot P(\boldsymbol{\theta}|\mathcal{S}_{\text{train}}) \cdot d\boldsymbol{\theta},$$

- MCMC methods generate weight sets for a given network architecture. The bayesian approach avoids overfitting the training data regardless of the network complexity.

3.- Classification approaches IV: Stacked generalization

Stacked generalization or stacking consists in combining several different classifiers into a unique class prediction by means of a new meta classifier.



4.- Evaluation

- ✗ All three approaches + stacking were evaluated using 10-fold cross validation.
- ✗ The consistent average error rate in all of them was 10.1% with a standard deviation of 4.9%
- ✗ Stacking did not introduce significant improvements with respect to the individual classifiers
- ✗ Errors are due to non separability of classes in the attribute hyperspace, bad quality data (outliers) in the photometric time series, bad period determinations and misclassifications in the original HIPPARCOS catalogue.

5.- Future steps I

- Establish a reliable procedure for period determination entirely without human supervision. SVO already has funding to tackle this task.
- Establish quality thresholds to avoid feeding the classifiers with low quality data (new experiments with OGLE data)
- Enlarge the input space (attributes) with as many diagnostics as possible (new photometric colours, spectral lines, metallicity indicators, spectral types...) in order to fully exploit the potential of Virtual Observatories and data mining techniques.

5.- Future steps II

- Enlarge the output space to cover as many variability types as possible including:
  - Class refinements (subclasify population I/II Cepheids, distinguish ELL and EW systems, Semirregulars...)
  - underpopulated classes in HIPPARCOS such as BY Dra, RS CVn, GCAS, etc.
- new variability types not included in HIPPARCOS (e.g. γ Doradus, WDs, solar type oscillations, sdB...)
- irregular, non periodic behaviour (flares, outbursts, novae,...), wavelet analysis.