STraDiWA: A simulation environment for astronomical transient discovery

RELATORI:
Prof. Giuseppe Longo
Prof. Massimo Brescia

CANDIDATO:
Marianna Annunziatella
Matr. N91/11
Synoptic surveys (early XXI century)

- Past surveys: Microlensing projects (OGLE, MACHO) have already collected millions of light curves.

- Currently operating: CRTS, PTF, VST, collect data streams of ~ 0.1 TB /night and detect ~ $10^{-10}$ transients/night.

- Forthcoming: DES, GAIA, LSST, EUCLID, will move us to the PB regime and will detect up to ~ $10^6 - 10^7$ transients event/night.
Time Domain Astronomy: a new axis of the OPS

Data sets many orders of magnitude larger, more complex, and more homogeneous than in the past.

All the observable quantities form the Observable Parameter Space (OPS).

Astronomical discovery comes from expansion or better sampling of OPS.

The advent of the synoptic surveys has meant that a new region of the OPS, the Time Domain, has been opened and extensively explored in the recent years.
Why the transients?

Variable objects are important in many astrophysical fields. For example:

• They are useful to better understand stellar structure and evolution.

• Many of them are distance indicators.

• We can find unknown types of objects and/or phenomena.

This thesis is finalized to the study and optimization – based on simulations - of Machine learning algorithms for real time variable object classification.
Part I: The SIMULATION OF the DATA STREAM
Data mining & Exploration Tool

Inspired by human brain features: high-parallel data flow, generalization, robustness, self-organization, pruning, associative memory, incremental learning, genetic evolution.

It is a web application for data mining experiments, based on WEB 2.0 technology.

![Diagram of data mining process]

- Multi Layer Perceptron trained by:
  - Back Propagation
  - Quasi Newton
  - Genetic Algorithm
- Support Vector Machines
- Genetic Algorithms
- Self Organizing Feature Maps
- K-Means
- Multi-layer Clustering
- Principal Probabilistic Surfaces
- Bayesian Networks
- Random Decision Forest
- MLP with Levenberg-Marquardt

next …
In order to be as realistic as possible, a simulation has to take into account as many as possible relevant factors:

- Survey strategy (how many bands, how deep).
- Sampling mode.
- Observing site conditions.
- Instrument setup.
- Realistic distribution of stars and galaxies
- Realistic distribution of pre-modeled variable objects.
**Starting Tools**

**STUFF**: Simulation of background galaxies.

**SkyMaker**: Produces an image starting from a catalog of galaxies, produced by STUFF, adding a RANDOM field of stars.

**Software for catalog extraction**.
Setup phase

- Number of bands and magnitude limits.
- Sampling: even or uneven.
- Observing Condition: the seeing FWHM can be assigned for each epoch or can be extracted randomly within a given range.
- Type and distribution of variable objects.
Stuff: creation of the static sky

Stuff generates a catalog of galaxies which take into account:

• Cosmological models.
• Luminosity distribution of the sources.
• Redshift distribution of the sources.
• Color distribution of the sources.
SkyMaker: Image production

The simulation of optics and detector, as well as the image rendering are carried out by SkyMaker.

SkyMaker works through several steps:

- PSF (Point Spread Function) modeling.
- Source modeling.
- Adding a uniform sky background.
- Applying Poissonian photon white noise and Gaussian read-out noise of the detector to the image.
We focused on two types of variable objects:

- Classical Cepheids, which show a periodic light curve.
- Type Ia Supernovae, which show a steep rise to the maximum phase.
- Random variable objects to preliminary simulate eruptive AGN.
Part II
OBJECT EXTRACTION
Comparison of source extraction software - I

We chose to make a comparison between SExtractor + PSFEx (Bertin & Arnouts 1996, Bertin 2011) and DAOPHOT + ALLSTAR (Stetson 1987, Stetson 1994).

We evaluated the following aspects:

- Completeness.
- Purity.
- Accuracy of photometry.
- Accuracy of the determination of centroids.

Comparison of object extraction software - II

The images used for the comparison was simulated using the characteristic of the VST optics, and the using $\frac{1}{4}$ of the size of the camera. We used an exposure time of 1500s and set the magnitude limits between 14 and 26 magnitude and the seeing at 0.7, an average value in Cerro Paranal. All the following comparisons are referred to a B-band image.

Stamp of a region of the simulated image in B band.
Completeness of the catalog - I

Since DAOPHOT is designed to perform stellar photometry, we compare the magnitude limit of the catalog by dividing the sources in stars and galaxies.

We define the limit in magnitude of the catalog (completeness) as the magnitude value below which $N_{\text{detected}}/N_{\text{input}}$ is less than 90%.
Source classification: SExtractor & PSFEx

Cross: Stars

Points: Galaxies
Cross: Stars

Points: Galaxies
We define as purity the ratio between stars correctly classified and detected stars ($S_{\text{classified}}/S_{\text{detected}}$). In analogy with the completeness an acceptable value for the purity is 0.9.
Photometric measurements can be derived from aperture photometry (top panels) or PSF fitting photometry (bottom panels).

Left panels: results obtained with DAOPHOT
Right panels: results obtained with SExtractor.

PSF fitting photometry provides more accurate measurements for the magnitude of the sources.
Comparison of object extraction software - Results

• SExtractor & PSFEx can detect source at one magnitude fainter than DAOPHOT. By using SExtractor however, we can maximize the detection of both point-like and extended sources.

• Down to their limit in magnitude, both DAOPHOT & ALLSTAR and SExtractor & PSFEx can classify stars with a purity greater than 0.9.

• Both DAOPHOT & ALLSTAR and SExtractor & PSFEx can produce accurate determination of magnitude and centroids of the objects.
We focused on two types of variable objects:

- **Classical Cepheids**, which show a periodic light curve.
- **Type Ia Supernovae**, which show a steep rise to the maximum phase.
- **Random variable objects** to preliminary simulate AGN.
Semantic tree of astronomical variables and transients

Credit: L. Eyer & N. Molnari (10/2007)
Semantic tree of astronomical variables and transients

Credit: L. Eyer & N. Mowlavi (10/2007)
Semantic tree of astronomical variables and transients

Credit: L. Eyer & N. Mowlavi (10/2007)
Semantic tree of astronomical variables and transients

Credit: L. Eyer & N. Mowlavi (10/2007)
Semantic tree of astronomical variables and transients

Variability Tree

Asteroids
- Rotation
- Eclipse

Stars
- Extrinsic
  - Asteroid occultation
  - Planetary transits
  - Eclipsing binary
  - Micro-lensing
  - Rotation
- Intrinsic
  - Eruptive
    - Cataclysmic
  - Pulsation
  - Secular

AGN

Credit: L. Eyer & N. Mowlavi (10/2007)
Semantic tree of astronomical variables and transients

Credit: L. Eyer & N. Mowlavi (10/2007)
Classical Cepheids have a periodic lightcurve, so their model is relatively simple.

As model for type Ia Supernovae we used an analytical function, used in Contardo, Leibundgut, and Vacca 2000:

\[
m(t) = \frac{f_0 + \gamma(t - t_0) + g_0 e^{\frac{(t-t_0)^2}{2\sigma_0^2}} + g_1 e^{\frac{(t-t_1)^2}{2\sigma_1^2}}}{(1 - e^{\frac{t-t_f}{\tau}})}
\]
Part III
REAL TIME CLASSIFICATION
PRELIMINARY RESULTS
Classification

A reliable classification of this type of objects is needed, especially to guarantee follow-up observations for short-lived transients.

Classification of variable object must ensure:

• a high completeness;
• a low contamination.

Furthermore, classification of variable objects must be as near real-time as possible, hence given the data stream, we need for machine learning methods.
Problems

• How to characterize variable objects (light curves, other statistical indicators)?

• Knowledge base built on the data themselves or rather on simulated ones?

• How to solve the computational challenge?

• How to find the unknown? Throwing away all the known (classification) or searching for intrinsic partitions of the Parameter Space (clustering)?
Problems

• How to characterize variable objects (light curves, other statistical indicators)?

• Knowledge base built on the data themselves or rather on simulated ones?

• How to solve the computational challenge?

• How to find the unknown? Throwing away all the known (classification) or searching for intrinsic partitions of the Parameter Space (clustering)?
Problems

• How to characterize variable objects (light curves, other statistical indicators)?

• Knowledge base built on the data themselves or rather on simulated ones?

• How to solve the computational challenge?

• How to find the unknown? Throwing away all the known (classification) or searching for intrinsic partitions of the Parameter Space (clustering)?
Problems

• How to characterize variable objects (light curves, other statistical indicators)?

• Knowledge base built on the data themselves or rather on simulated ones?

• How to solve the computational challenge?

• How to find the unknown? Throwing away all the known (classification) or searching for intrinsic partitions of the Parameter Space (clustering)?
Problems

• How to characterize variable objects (light curves, other statistical indicators)?

• Knowledge base built on the data themselves or rather on simulated ones?

• How to solve the computational challenge?

• How to find the unknown? Throwing away all the known (classification) or searching for intrinsic partitions of the Parameter Space (clustering)?
A Hierarchical approach to classification

Different types of classifiers perform better for some event classes than for the others.

We propose to find which is the best classifier for a particular type of variable object.

Our approach has the typical decision tree structure and aims at a classification which becomes finer and finer as we go to higher level of branching.
Data mining & Exploration Tool

Inspired by human brain features: high-parallel data flow, generalization, robustness, self-organization, pruning, associative memory, incremental learning, genetic evolution.

It is a web application for data mining experiments, based on WEB 2.0 technology.

Multi Layer Perceptron trained by:
- Back Propagation
- Quasi Newton
- Genetic Algorithm

Support Vector Machines
Genetic Algorithms
Self Organizing Feature Maps
K-Means
Multi-layer Clustering
Principal Probabilistic Surfaces

Bayesian Networks
Random Decision Forest
MLP with Levenberg-Marquardt

next ...
Classification
Regression
Clustering
Feature Extraction
First classification test: MLP-QNA

 MLP-QNA is a MLP trained by a Quasi Newton rule (QNA).

\[ \nabla^2 E(w_k) d_k = -\nabla E(w_k) \]

Hessian approx. (QNA)

This algorithm has already given the best results with other astrophysical problems.

(Brescia et al. 2012, arXiv:1110.214444)
Features choice

Since we can not use the information available for all the epochs we have to specify the number of the epochs and how they must be chosen:

- $N$ random epoch different for each object.
- $N$ random epoch equal for each object.
- $N$ epoch equally spaced.

As features for the classifier we used the pairs of parameters $(m_i, t_i)$ for ten epochs randomly chosen for each object.
## Statistical Results

Evaluation criteria:
- Accuracy (CA).
- Purity (CO).
- Contamination (CN).

Increasing the size of the sample there is a no significant improvement of the accuracy, while there is large increase of the contamination.

<table>
<thead>
<tr>
<th>Object Number</th>
<th>Object Number</th>
<th>TRAINING</th>
<th>TEST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not Variable</td>
<td>Variable</td>
<td>CA %</td>
<td>CO %</td>
</tr>
<tr>
<td>SIM1</td>
<td>1187</td>
<td>99</td>
<td>96</td>
</tr>
<tr>
<td>SIM2</td>
<td>3319</td>
<td>99</td>
<td>98</td>
</tr>
<tr>
<td>SIM3</td>
<td>6338</td>
<td>96</td>
<td>97</td>
</tr>
<tr>
<td>SIM4</td>
<td>6362</td>
<td>08</td>
<td>96</td>
</tr>
</tbody>
</table>
Conclusions

• We presented a modular simulation framework, based on reliable software tools, integrated in a workflow specialized to variable object realistic representation.

• We have successfully modeled and simulated a preliminary subset of variable objects, populating realistic multi-band images, as observed by the VST and OmegaCAM instruments.

• We performed a deeper performance analysis and comparison between source extraction software in order to enhance and optimize their capability to detect transients (Annunziatella et al., 2012).

• We presented here some preliminary results obtained by a set of experiments, based on the MLPQNA.
Forthcoming developments

In the next future we plan to:

- investigate the use of statistical indicators as features for the classifiers, in order to optimize the classification between the variable and non variable objects.

- Add other types of variable objects to the simulation, in particular expanding the category of SNe and pulsating variables, in order to go down to the next levels of the branching.
Grazie per l’attenzione