Mem. S.A.It. Suppl. Vol. 13, 56
© SAIt 2009

Memorie della
Supplementi

Astrophysics in S.Co.P.E.

M. Brescia[1], S. Cavuoti[2], G. d'Angelo[2], R. D'Abrusco[2], C. Donalek[3], N. Deniskina[2],
O. Laurino[2], and G. Longo[2,1,4]

[1] INAF - Osservatorio Astronomico Capodimonte, via Moiariello 16, 80131, Napoli, Italy
[2] Department of Physical Sciences - University Federico II, Naples, Italy
e-mail: longo@na.infn.it
[3] Department of Astronomy, California Institute of Technology, Pasadena, CA, USA
[4] INFN - Napoli Unit, via Cintia 9, 80126, Napoli, Italy

Abstract. S.Co.P.E. is one of the four projects funded by the Italian Government in order to provide Southern Italy with a distributed computing infrastructure for fundamental science. Besided being aimed at building the infrastructure, S.Co.P.E. is also actively pursuing research in several areas among which astrophysics and observational cosmology. We shortly summarize the most significant results obtained in the first two years of the project and related to the development of middleware and Data Mining tools for the Virtual Observatory.

Key words. distributed computing, cosmology

1. Introduction

S.Co.P.E. is a general purpose GRID infrastructure of the University Federico II in Naples funded through the Italian National Plan (PON) by the Italian Government to support both fundamental research and small/medium size companies. The infrastructure has been conceived as a metropolitan GRID, embedding different (and in some cases pre-existing) and heterogeneous computing centers each with its specific vocation: high energy physics, astrophysics, bioinformatics, chemistry and material sciences, electric engineering, social sciences. Its intrinsically multi-disciplinary nature renders the S.Co.P.E. an ideal test bed for innovative middleware solutions and for interoperable tools and applications finely tuned on the needs of a distributed computing environment. In what follows we shall shortly outline the main activities in the fields of astrophysics and observational cosmology and, in particular, we shall focus on: i) the ongoing efforts aimed at integrating the S.Co.P.E. GRID (hereafter SG) with the international Virtual Observatory (Sect.2), and ii) the implementation in the SG of the data mining (DM) VO-Neural package (Sect.3) which is developed in the framework of a collaboration with the Dept. of Astronomy at Caltech. In Sect. 4 we shortly outline a template scientific application and, finally, in Sect. 5, we outline some future developments.

2. The VOb and the GRID

The Virtual Observatory (VOb) is an international effort coordinated through the International Virtual Observatory Alliance

Send offprint requests to: G. Longo

IVOA; URL.1 (2000) aimed at: i) federating and making interoperable all astronomical data archives produced by both ground based and space borne instruments; ii) deploying a new generation of science applications or tools which use VOb protocols for exploratory data analysis and for the extraction of knowledge from massive data sets. The VOb is inherently distributed: data collections remain with their providers and are accessed through standard interfaces. The access to the data takes place through a registry which contains information about data sets, archives, catalogs, surveys, and computational services that can be accessed through VOb interfaces Hanish & De Young (2008). While the federation and fusion of heterogeneous data archives and the implementation of flexible data reduction and data analysis tools have been widely addressed and, at least in their fundamental aspects, solved, the possibility to access large distributed computing facilities to perform computing intensive tasks has not yet been satisfactorily answered.

One problem to be solved is the conflict existing between the VOb and the GRID security procedures: most users of a specific Virtual Organization (VO) do not possess the personal certificates which are requested to access the GRID or, even when they do have a personal certificate, the computing GRID which they need does not recognize their own certification authority.

In the framework of the VONeural project (Sect. 3) and in order to make our Data Mining (DM) tools accessible to the wider community, we implemented and tested a general purpose interface between the UK-ASTROGRID hereafter AG; URL.4 (2000) and the SG.

2.1. GRID-Launcher v.1.0

The UK based ASTROGRID is one of the most robust astronomical Virtual Organizations so far implemented and represents a good ground for testing innovative solutions. The main problem we had to face was the fact that most users which are recognized by the AG User Authentication Service do not possess a personal GRID certificate and cannot therefore access distributed computing resources. This problem can be at least in part circumvented by offering the applications as web services to be consumed on the Grid via a service certificate (or ”robot” certificates). At the time GRID-Launcher v.1.0 was developed, this option had not yet been formally accepted by the EGEE-2 boards and we were obliged to implement a test version which makes use of a personal GRID certificate (signed by the INFN-GRID CA) which is recognized by the S.Co.P.E. GRID.

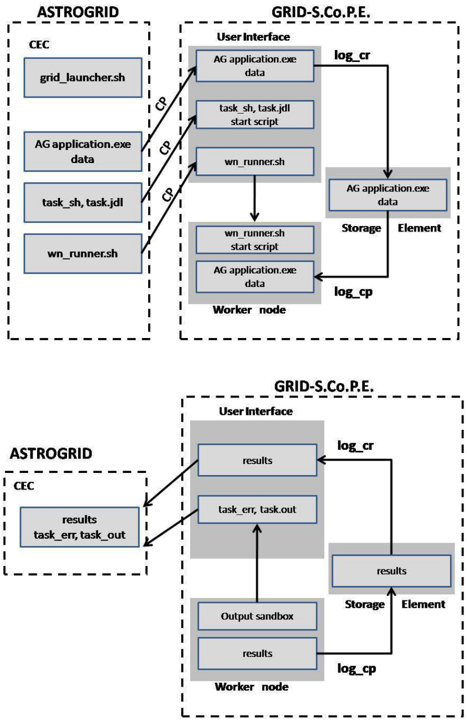

Fig. 1. Grid Launcher v.1.0 workflows for input and output. UI: user interface; RB: resource broker; SE: storage element; CE: computing element; WN: working node. Upper panel: imput flow; lower panel: output flow.

In a very schematic way, GRID-Launcher works as it is summarized in Fig. 1:

- It takes the user input from the User Interface of the ASTROGRID Desktop, collects all files, tabs and programs needed and generates automatically three scripts:

task.sh, *task.jdl* and *wn_runner.sh* to be executed on the GRID;
- it wraps them in an archive and sends it to the GRID UI (authentication takes place with public keys exchange);
- the UI unpacks them, copies the data to the Storage Element (SE), copies *wn_runner.sh* to the WN's, starts *task.sh* and *task.jdl*;
- *wn_runner.sh* starts on the WNs, takes the data from SE, starts the application and puts the results on the SE. The GRID generates automatically two output files *task.err* and *task.out* and sends them to the UI using the Output SandBox.
- GRID-launcher periodically checks the status of job and when it ends, it moves the results from the UI to the ASTROGRID machine. GRID-launcher receives the data archive, unpacks them and puts the results into the AG Myspace (VO-Space).

So far, GRID-Launcher v.1.0 has been implemented and tested on an handful of applications: VO-Neural_MLP & VO-Neural_SVM (cf. Sect. 3), Sextractor Bertin & Arnouts (1996) & SWarp URL.8 (2000).

3. VO-Neural

As it was mentioned above, in the last decade many national cf. URL.2 (2000) and international cf. URL.3 (2000) projects have solved many problems related to the federation of heterogeneous data sets while much remains still to be done for what tools and user interfaces are concerned. One of the main issues to be solved is the implementation of scalable and user friendly data mining tools capable to deal with the huge VOb data sets.

VO-Neural is a data mining framework capable to work on massive ($>$ 1 TB) data sets (catalogues) in a distributed computing environment matching the IVOA standards and requirements. VO-Neural is the evolution of the AstroNeural Tagliaferri et al. (2003) project which was started in 1994, as a collaboration between the Department of Mathematics and Applications at the University of Salerno and the Astronomical Observatory of Capodimonte-INAF, and is currently under continuous evolution. VO-Neural allows to extract from large datasets information useful to determine patterns, relationships, similarities and regularities in the space of parameters, and to identify outlayers. In its final version, it will be accessible both as a web application and through the AG Desktop and will offer main elaborative features like exploratory data analysis, data prediction and ancillary functionality like fine tuning, visual exploration of the main characteristics of the datasets, etc.. Besides offering the possibility to use the individual routines to perform specific tasks, VO-Neural will provide the user with a complete framework to write his own customized programs. In the next two paragraphs we shortly outline the main features of two supervised clustering models already included in the package which have already been used on the GRID-S.Co.P.E. for specific science applications.

3.0.1. *VONeural_MLP*

VONeural_MLP is an implementation of a standard Multi Layer Perceptron based on the FANN (Fast Artificial Neural Networks) Library URL.6 (2000), written in C++ Skordovski (2008). The algorithm known as Multi Layer Perceptron (MLP) is based on the concept of perceptron and the method of learning is based on gradient-descent method that allows to find a local minimum of a function in a space with N dimensions. The weights associated to the connections between the layers of neurons are initialized at small and random values, and then the MLP applies the learning rule using part of the template patterns. Once convergence has been achieved and a validation procedure has been applied in order to avoid overfitting, the performances of the network are evaluated on a disjoint test set extracted from the template patterns. The resulting network is then applied to the original data.

3.0.2. *VONeural_SVM*

VONeural_SVM is an implementation of the Support Vector Machines Russo (2007); Cavuoti (2008) based on the LIBSVM library URL.7 (2000). Support Vector Machines perform classification of records into classes by first mapping the data into an higher dimensionality and then using a set of template vectors (targets) to find in this new space a separation hyperplane with the largest margin. Withouth entering into details Boser et al. (1992), we shall just remember that, in the case of the C-SVC implemented with the RBF (Radial Basis Functions) kernel, the position of this hyperplane depends on two parameters (C and γ) which cannot be estimated in advance but need to be evaluated by finding the maximum in a grid of values which is usually defined by letting C and γ vary as $C = 2^{-5}, 2^{-3}, ..., 2^{15}$ and $\gamma = 2^{-15}, 2^{-13}, ..., 2^{3}$. Due to their computational weight, and to the need to run many iterations for different pairs of the two parameters, SVM are ideally suited for being used on the GRID.

4. An application to the classification of AGN in the SDSS

The astronomical community is used to perform DM tasks in a sort of ”hidden” way (cf. the case of specific objects selection in a color-color diagram) but it has not yet become familiar with the potentialities of more advanced tools such as those described here. This is mainly due to the fact that these tools are often everything but user friendly and require an in depth understanding of the (often complex) theory laying behind them; a complexity which often discourages potential users. Therefore, a crucial aspect of the project is the application to challenging problems capable to exemplify the new science which will emerge from the adoption of a less conservative approach to the analysis of the data. Two science cases, namely the evaluation of photometric redshifts (a regression and classification problem based on the use of *VONeural_MLP*) and the selection of candidate quasars in the Sloan Digital Sky Survey cf. Stoughton et al. (2002) (based on the use of unsupervised clustering algorithms and agglomerative clustering) have already been published in the literature D'Abrusco et al. (2007, 2008). We shall therefore focus on the application of *VONeural_SVM* to the classification of AGNs.

The classification of AGN is usually performed on their overall spectral distribution using some spectroscopic indicators (equivalent linewidths, FWHM of specific lines or lines flux ratios) and diagnostic diagrams (usually called BPT) which are difficult and time consuming to derive. In this diagrams AGN and not-AGN are empirically separated by some lines derived either from the theory or from empirical laws such as those derived by Kewley et al. (2006); Kauffman et al. (2003); Heckman (1980). A reliable and accurate AGN classificator based on photometric features only, would allow to save precious telescope time and enable several studies based on statistically significant samples of objects. We therefore used a supervised clustering of the photometric data exploiting the information contained in a spectroscopic Base of Knowledge (BoK) derived from available catalogues. We wish to stress that since neural networks have no power of extrapolation all the biases in the BoK are reproduced and therefore the BoK needs to be as complete and bias-free as possible. As classification tools, we used both the MLP and, due to the intrinsically binary nature of the problem (AGN against non-AGN, Seyfert 1 against Seyfert 2, etc) also the SVM.

The BoK was obtained from the fusion of two catalogues.

- Sorrentino et al (2006) separated objects into Seyfert 1, Seyfert 2 and ”Not AGN” using the Kewley's lines Kewley et al. (2006);
- a catalogue derived by us from the SDSS spectroscopic archive using the criteria introduced by Kauffman et al. (2003) in which objects are classified as AGN, not AGN, and ”mixed”. The Mix and Pure AGN zone were further divided into

Table 1. Summary of the results of supervised classification experiments performed using both *VONeural_MLP* and *VONeural_SVM*.

experiment	BoK	algorithm	efficiency	completeness
AGN vs Mix	BPT plot + Kewley line	MLP	76%	54%
	BPT plot + Kewley line	SVM	74%	55%
Type 1 vs 2	BPT plot + Kewley line	MLP	95%	~ 100%
	BPT plot + Kewley line	SVM	82%	98%
Seyfert vs LINER	BPT plot + Hecman & Kewley lines	MLP	80%	92%
	BPT plot + Kewley line	SVM	78%	89%

Seyfert and LINERs by using the Heckman lineHeckman (1980).

We made three experiments using both the MLP and SVM, and for all of them we used the same set of features (for a definition refer to the SDSS specifications) extracted from the SDSS database: *petroR50_u*, *petroR50_g*, *petroR50_r*, *petroR50_i*, *petroR50_z*, *concentration_index_r*, *fibermag_r*, $(u-g)dered$, $(g-r)dered$, $(r-i)dered$, $(i-z)dered$, *dered_r*, together with the photometric redshift in D'Abrusco et al. (2008). We performed three types of classification experiments: AGN vs Mix, Type1 vs Type2, Seyfert vs LINER. The experiments with SVM were performed on the SG using 110 worker nodes. In order to test the interoperability of the four PON projects, the 110 nodes were taken from the Napoli, Catania and Cagliari PON locations. Results are summarized in Table 4 and, as it can be seen, the use of machine learning tools allows to reach performances which in some cases (e.g. Type 1 vs 2 with MLP's) cannot by any means be achieved with more traditional tools. A more detailed discussion of the results will be presented in (Cavuoti, d'Abrusco & Longo, 2008, in preparation).

5. Future developments

We plan to continue the development of VO-Neural and to offer it as a web application. More in detail, we plan to deploy a general purpose GRID-Launcher interface capable to launch any "command line" program through a "robot certificate" (GRID-Launcher 2.0).

At the moment we are engineering the package in order to increase its flexibility and capability to adapt to a distributed computing environment. We are also implementing parallel versions of some tools which are particularly demanding in terms of computing time. We also plan to integrate, within the VO-Neural interface existing statistical software (such as, for instance, the VO-STAT web application URL.9 (2000)), in order to ensure proper statistical tools for exploratory data analysis and for the evaluation of the results. The status of the project can be monitored at the URL: http://voneural.na.infn.it/ .

Acknowledgements. The authors wish to thank M. Paolillo and E. de Filippis for many useful discussions. The work was funded through the Euro VO-Tech project and the MUR funded PON-S.Co.P.E.

References

Bertin E. & Arnouts S. 1996, A& AS, 117, 393.

Boser B. E., Guyon I. & Vapnik V. 1992,in Proc. of the Fifth Annual Workshop on Computational Learning Theory, 144, ACM Press.

Cavuoti S. 2008, M.Sc. Thesis, University of Napoli Federico II.

D'Abrusco R., Staiano A., Longo G., Brescia M., De Filippis E., Paolillo M., Tagliaferri R. 2007, ApJ, 663, 752.

D'Abrusco R., Longo G. & Walton N.A. 2008, astro-ph/0805.0156v1.

Hanish R. & de Young D. 2008, in The National Virtual Observatory Book ASP

Conference Series, Vol. 382, M. J. Graham M. J. Fitzpatrick, & T. A. McGlynn eds., 1.

Heckman T.M. 1980, A& A, 87, 182.

Kauffman G. et al. 2003, MNRAS, 346, 1055.

Kewley L.J. et al. 2006, MNRAS, 372, 961.

Skordovski B. 2008, M.Sc. Thesis, University of Napoli Federico II.

Russo V. 2007, M.Sc. Thesis, University of Napoli Federico II.

Stoughton C., Lupton R. H., Bernardi M. et al. 2002, AJ, 123, 485.

Tagliaferri R., Longo G., Milano L., Acernese F., et al. 2003, Neural Networks, 16, 297.

Sorrentino G., Radovich M. & Rifatto A. 2006, A & A 451, 809.

URL.1: IVOA: http://www.ivoa.org/

URL.2: NVO: http://www.nvo.org/

URL.3: Euro-VO: http://www.eurovo.org/

URL.4: http://www.astrogrid.uk/

URL.5: http://voneural.na.infn.it/

URL.6: http://leenissen.dk/fann/

URL.7: http://www.csie.ntu.edu.tw/ cjlin/libsvm/

URL.8: SWarp User manual , E. Bertin at: http://terapix.iap.fr/rubrique.php?id_rubrique=49

URL.9: VO-STAT, http://astrostatistics.psu.edu/