DAME: A WEB ORIENTED INFRASTRUCTURE FOR SCIENTIFIC DATA MINING AND EXPLORATION

S. Cavuoti*

Department of Physics, Universita' degli Studi Federico II,
Via Cintia 26, 80125 Napoli, Italy
City, State ZIP/Zone, Country
**E-mail: s.cavuoti@gmail.com*
http://dame.dsf.unina.it

M. Brescia

INAF - Osservatorio Astronomico di Capodimonte,
Via Moiariello 16, 80131 Napoli, Italy E-mail: brescia@na.astro.it

G. Longo, M. Garofalo AND A. Nocella

Department of Physics, Universita' degli Studi Federico II,
Via Cintia 26, 80125 Napoli, Italy
City, State ZIP/Zone, Country

Nowadays, many scientific areas share the same broad requirements of being able to deal with massive and distributed datasets while, when possible, being integrated with services and applications. In order to solve the growing gap between the incremental generation of data and our understanding of it, it is required to know how to access, retrieve, analyze, mine and integrate data from disparate sources. One of the fundamental aspects of any new generation of data mining software tool or package which really wants to become a service for the community is the possibility to use it within complex workflows which each user can fine tune in order to match the specific demands of his scientific goal. These workflows need often to access different resources (data, providers, computing facilities and packages) and require a strict interoperability on (at least) the client side. The project DAME (DAta Mining & Exploration) arises from these requirements by providing a distributed WEB-based data mining infrastructure specialized on Massive Data Sets exploration with Soft Computing and machine learning methods. It results as a multi-disciplinary platform-independent tool perfectly compliant with modern KDD (Knowledge Discovery in Databases) requirements and Information & Communication Technology trends.

Keywords: Astroinformatics; data mining, knowledge discovery in database; machine learning; virtual observatory.

1. Introduction

In almost all fields, modern technology allows to capture and store huge quantities of heterogeneous and complex data often consisting of hundreds of features for each record and requiring complex metadata structures to be understood.. A need has therefore emerged for a new generation of software tools, largely automatic, scalable and highly reliable. Strictly speaking, Knowledge Discovery in Databases (KDD) is about algorithms for inferring knowledge from data and ways of validating the obtained results, as well as about running them on infrastructures capable to match the computational demands. In practice, whenever there is too much data or, more generally, a representation in more than 5 dimensions,[1] DAME copes with this problem by implementing the KDD models under the S.Co.P.E. (Scientific Cooperation for Pluri-disciplinary Experiments) grid[9] and by taking into account the fact that background knowledge can make it possible to reduce the amount of data that needs to be processed by adopting a learning rule based on the fact that in many cases most of the attributes turn out to be irrelevant when background knowledge is taken into account.[2]

Data Mining requires a lengthy fine tuning phase which is often not easily justifiable to the eyes of a non experienced user.

Such complexity is one of the main explanations for the gap still existing between the new methodologies and the huge community of potential users which fail to adopt them. In order to be effective, in fact, Knowledge Discovery in Databases requires a good understanding of the mathematics underlying the methods, of the computing infrastructures and of the complex workflows which need to be implemented, and most users (even in the scientific community) are usually not willing to make the effort to understand the process and prefer to recur to traditional approaches which are far less powerful but much more user friendly.[3]

DAME, by making use of the Web application paradigm, of extensive and user friendly documentation and of a sample of documented and realistic use cases, represents a first attempt to bring the KDD models to the user hiding most of their complexity behind a well designed infrastructure. In this paper we describe the prototype (alpha release) version of the DAME (Data Mining and Exploration) web application, nowadays available under final commissioning which addresses many of the above issues and aims at providing the scientific community with a user friendly and powerful data mining platform.

2. General Description

DAME was initially conceived to work on astrophysical Massive Data Sets data as a tool offered to the community and aimed at solving some of the above quoted problems by offering a completely transparent architecture, a user friendly interface and the possibility to access a distributed computing infrastructure.

DAME starts from a taxonomy of data mining methods (hereinafter called functionalities) and collects a set of machine learning algorithms (hereinafter called models) that can be associated to one or more functionalities depending of the specific problem domain. This association "functionality-model" represents what in the following we shall refer to as "experiment domain".

At a lower level, any experiment launched on the DAME framework, externally configurable through dynamical interactive web pages, is treated in a standard way, making completely transparent to the user the specific computing infrastructure used and the specific data format given as input. Dimensional reduction, classification, regression, prediction, clustering, filtering, are examples of functionalities belonging to the data mining conceptual domain, in which the various methods (models and algorithms) can be applied to explore data under a particular aspect, connected to the associated functionality scope. In its first implementation the infrastructure prototype has been focused on the classification, regression and clustering functionalities.

A special care was devoted in each single phase of the design and development to produce a complete and exhaustive documentation both technical and user oriented. . As The concept of "distributed archives" is familiar to most scientists. In the astronomical domain, the leap forward was the possibility to organize through the Virtual Observatory (hereafter VObs)[4] the data repositories to allow efficient, transparent and uniform access. In other words, the VObs was intended to be a paradigm to use multiple archives of astronomical data in an interoperating, integrated and logically centralized way, so to be able to "observe" and analyze a virtual sky by position, wavelength and time.[14] The link between data mining applications and the VObs data repositories is currently still under discussion since it requires (among the other things) the harmonization of many recent achievements in the fields of VObs, grid, cloud, HPC (High Performance Computing), and Knowledge Discovery in Databases. DAME was conceived to provide the VObs with an extensible, integrated environment for data mining and exploration. In order to do so, DAME had to support the VObs standards and formats, especially for data interoperability and to abstract the application deployment and execution, so to provide the VObs with a general purpose computing platform taking advantage of the modern technologies.

In order to gradually accomplish such requirements, DAME intended to give the possibility to remotely interact with data archives and data mining applications via a simple web browser.[5] Thus, with web applications, a remote user does not require to install program clients on his desktop, having the possibility to collect, retrieve, visualize and organize data, configure and execute mining applications through the web browser and in an asynchronous way. An added value of such approach being the fact that the user does not need to directly access large computing and storage power facilities. Connected with the dichotomy between supervised and unsupervised learning, in the DAME data mining infrastructure, the choice of any machine learning model is always accompanied by the functionality domain. In what follows we shall therefore adopt the following terminology:

- Data mining model: any of the data mining models integrated in the DAME suite. It can be either a supervised machine learning algorithm or an unsupervised one, depending on the available Base of Knowledge (BoK, i.e. the set of training or template cases available) and the scientific target of the user experiment;
- Functionality: one of the functional domains in which the user wants to explore the available data (for example, regression, classification or clustering). The choice of the functionality target can limit the choice of the data mining model;
- Experiment: it is the scientific pipeline (including optional preprocessing or preparation of data) and includes the choice of a combination of data mining model and a functionality;
- Use Case: for each data mining model, different running cases are exposed to the user . These can be executed singularly or in a prefixed sequence. Being the models derived from the machine learning paradigm,[6] each has training, test, validation and run use cases, in order to, respectively, perform learning, verification, validation and execution phases. In most models there is also the "full" use case, that automatically executes all listed cases as a sequence.

The functionalities and models, already available and under deployment in the DAME web application, are the following:

- Classification and Regression (Supervised): three different versions

of the classical MultiLayer Perceptron (MLP) trained, respectively by Back Propagation, Genetic Algorithms and Quasi Newton Algorithms; Support Vector Machines (SVM);
- Dimensional Reduction (unsupervised): Principal Probabilistic Surfaces (PPS);
- Clustering and Image Segmentation (unsupervised): Self Organizing Maps (SOM);

3. The Architecture

The DAME design architecture is implemented following the standard LAR (Layered Application Architecture) strategy, which leads to a software system based on a layered logical structure, where different layers communicate with each other via simple and well-defined rules:

- Data Access Layer (DAL): the persistent data management layer, responsible of the data archiving system, including consistency and reliability maintenance.
- Business Logic Layer (BLL): the core of the system, responsible of the management of all services and applications implemented in the infrastructure, including information flow control and supervision.
- User Interface (UI): responsible of the interaction mechanisms between the BLL and the users, including data and command I/O and views rendering.

A direct implication of the LAR strategy adopted in DAME is the Rich Internet Application (RIA),[7] consisting in applications having traditional interaction and interface features of computer programs but usable via simple web browsers, i.e. not needing any installation on user local desktop. RIAs are particularly efficient in terms of interaction and execution speed. By keeping this in mind, the main concepts behind the distributed data mining applications implemented in the DAME Suite are based on three issues:

- Virtual organization of data: extension of the already remarked basic feature of the VObs;
- Hardware resource-oriented: this is obtained by using computing infrastructures, like grid, which enable parallel processing of tasks, using idle capacity. The paradigm in this case is to obtain large numbers of instances running for short periods of time;

- Software service-oriented: this is the base of typical cloud computing paradigm.[8] The data mining applications implemented runs on top of virtual machines seen at the user level as services (specifically web services), standardized in terms of data management and working flow.

The complete Hardware infrastructure of the DAME Program, where the grid sub-architecture is provided by the S.Co.P.E. supercomputing facility,[139] is incorporated into the more general cloud scheme, including a network of multi-processor PCs and workstations, each of them internally dedicated to a specific function. The integrity of the system, including the grid public access, is guaranteed by a registration procedure, giving the possibility to access all facilities from just one account. In particular, a robot certificate is automatically handled by the DAME system to provide transparent access to the S.Co.P.E. grid resources.[10] Depending on the computing and storage power, requested by the job and by the processing load currently running on the network, an internal mechanism redirects the jobs to a job-queue in a pre-emptive scheduling scheme. The interaction with the infrastructure is completely asynchronous and a specialized software component (DR, DRiver) has the responsibility to store off-line job results in the user storage workspaces, that can be retrieved and downloaded in subsequent accesses. This hybrid architecture, renders it possible to execute simultaneous experiments that gathered all together, bring the best results. Even if the individual job is not parallelized, we obtain a running time improvement by reaching the limit value of the Amdahl's law (N):

$$\frac{1}{1 - P + \frac{P}{N}} \tag{1}$$

where P is the fraction of a program that can be made parallel (i.e. which can benefit from parallelization), and (1 - P) is the fraction that cannot be parallelized (remains serial), then the resulting maximum speed-up that can be achieved by using N processors is obtained by the law expressed above.

For instance, in the case of the AGN (Active Galactic Nucleus) classification experiment detailed in,[11],[12] each of the 110 jobs runs for about a week on a single processor. By exploiting the grid, the experiment running time can be reduced to about one week instead of more than 2 years.

The DAME software architecture is based on five main components: Front End (FE), Framework (FW), Registry and Data Base (REDB), Driver (DR) and Data Mining Models (DMM).

The Front End (FE) is the component directly interfacing the end user with the infrastructure (through the web browser). and it can be considered structured in two main parts (also addressable as a ”client” and a ”server”, even though both of them are resident on the remote side with respect of the user): the main GUI (Graphical User Interface) of the Suite and the internal interface (hereinafter Front End Server) with the inner infrastructure.

4. Conclusions and Future Developments

DAME is an evolving platform and new modules and specific workflows as well as additional features are continuously added. The modular architecture of DAME can also be exploited to build applications, finely tuned to specific needs. Examples available so far and accessible through the DAME website, being VOGCLUSTERS (Virtual Observatory Globular Clusters), a VObs web application aimed at collecting and make available all existing data on galactic globular clusters for data and text mining purposes, and NExt-II (Neural Extractor) for the segmentation of wide field astronomical images. The main product I a Data Mining Web Application suite avaialbe at address: http://dame.dsf.unina.it/beta_info.html

Moreover, it is foreseen the introduction of MPI (Message Passing interface) technology, by investigating its deployment on a multi-core platform, based on GPU+CUDA computing technique.[15] This could improve computing efficiency in data mining models, such as Genetic Algorithms, naturally implementable in a parallel way. In conclusion, we are confident that DAME may represent what is generally considered an important element of e-science: a stronger multi-disciplinary approach based on the mutual interaction and interoperability between different scientific and technological fields (nowadays defined as X-Informatics, such as Astro-Informatics).

Acknowledgments

AAll partners acknowledge the financial support of the Italian Ministry of Foreign Affairs for the Italy-USA bi-lateral grant ”Building an e-science Data Mining Infrastructure”, the European Union through the projects VO-Tech and VO-AIDA and the Ministry of University and Research through the PON S.Co.P.E. GRID project. SGD, AAM, and CD acknowledge a crucial support from the Fishbein Family Foundation, and a partial support from the NASA grant 08-AISR08-0085

References

1. S. Odenwald, "Cosmology in More Than 4 Dimensions", Astrophysics Workshop, N.R.L., 1987
2. G. Paliouras, "Scalability of Machine Learning Algorithms", M. Sc. Thesis, University of Manchester, 1993
3. T. Hey, S. Tansley, K. Tolle, "The Fourth Paradigm: Data-Intensive Scientific Discovery", Microsoft Research, ISBN-10: 0982544200, 2009
4. F. Genova, G. Rixon, F. Ochsenbein, C.G. Page, "Interoperability of archives in the VO", Proceedings of SPIE Conference Virtual Observatories, Alexander S. Szalay Editor, Vol. 4846, pp.20-26, 2002
5. M. Brescia, G. Longo, F. Pasian, "Mining Knowledge in Astrophysical Massive Data Sets", Nuclear Instruments and Methods in Physics Research, Section A, Vol. 623, Issue 2, pp. 845-849, Elsevier Science, ISSN 0168-9002, 2010
6. C.M. Bishop, "Pattern Recognition and Machine Learning", Springer ISBN 0-387-31073-8, 2006
7. A. Bozzon, et al., "Engineering Rich Internet Applications with a Model-Driven approach", ACM Transaction on the Web (TWEB), vol. 4, p. 7.1-7.47, 2010
8. J. Shende, "Service-Oriented Architecture and the Cloud", Cloud Expo, SYS-CON Media, http://cloudcomputing.sys-con.com/node/1524663, 2010
9. L. Merola, "The SCOPE Project", Proceedings of the Final Workshop of GRID Projects PON RICERCA 2000-2006, AVVISO 1575. Catania, Italy, 2008
10. N. Deniskina, et al., "GRID-Launcher v.1.0", Contributed to Data Centre Alliance Workshops: GRID and the Virtual Observatory, Memorie della Societa Astronomica Italiana, Vol.80, p.571, Munich, 2009
11. S. Cavuoti, et al., "Photometric AGN classification in the SDSS", Italian E-science 2008 Conference Naples, Italy, 2008
12. M. Brescia, et al., "Astrophysics in S.Co.P.E", Mem. S.A.It. Suppl. Vol 13, 56, 2009
13. M. Brescia, et al., "DAME: A Distributed Web Based Framework for Knowledge Discovery in Databases", 54th SAIT Conference Next Decade Prospective for the Italian Astronomy, Astronomical Observatory of Capodimonte, Napoli, (accepted in press), 2009
14. G. Fabbiano, D. Calzetti, C. Carilli, S. G. Djorgovski, et al., "Recommendations of the VAO-Science Council", arXiv:1006.2168v1 [astro-ph.IM], 2010
15. B. Maier, "High Performance Computing with CUDA", American Physical Society, Joint Fall 2009 Meeting of the Texas Sections of the APS, AAPT, and SPS Post deadline, abstract #C5.006, 2009