DAME: a distributed data mining infrastructure for e-science discoveries
http://voneural.na.infn.it/

M. Brescia¹, R. D’Abrusco², C. Donalek³, S.G. Djorgovski³, O. Laurino², G. Longo², A. Mahabal³,
S. Cavuoti², G. d’Angelo², M. Garofalo², A. Nocella²

1 - INAF — Osservatorio Astronomico di Capodimonte, Napoli, Italy; 2 - Department of Physics, University Federico II, Napoli, Italy;
3 - Department of Astronomy, California Institute of Technology, Pasadena, USA

In many e-science communities and environments Massive Data Sets (MDS) are gathered by means of heterogeneous techniques and stored in very diversified and often-incompatible data repositories. Moreover, it is needed to integrate services across distributed, heterogeneous, dynamic “virtual organizations” formed by resources available within a single enterprise and/or from external resource sharing and service provider relationships. The DAME project aims at creating a general purpose Data Mining and Exploration e-infrastructure capable to ensure integrated and asynchronous processing of data stored in such MDS.

The project consists of a data mining framework with powerful software tools capable to work on massive data sets, compliant with Virtual Observatory standards, in a distributed computing environment. The integration process can be technically challenging because of the need to achieve a specific quality of service when running on top of different native platforms. In these terms, the result of the DAME project effort is a service-oriented architecture, by using appropriate standards and incorporating Cloud/Grid paradigms and Web services, that will have as main target the integration of interdisciplinary distributed systems within and across organizational domains.

It is a web oriented and VO aware suite. In order to perform data mining and exploration experiments, we have considered a top-down strategy, starting from the taxonomy of data mining and research functionalities which are associated to specific algorithms and processing methods. In the first release, the suite offers tools for the following functionalities:

- Classification
- Regression
- Clustering

The above picture shows both already implemented models and next foreseen release evolutions.

Main features of the complete package:
- Object Oriented Programming & UML
- Internal standards and protocols (XML)
- Java language (generic for DMM)
- User/Session Registry DB (MySQL)
- Web-based User I/O (GWT-Ext)
- Web Application and Web Service Technology
- Plugin Modularity (easy to be integrated/modified)
- Hardware independent through platform driver
- Data conversion and manipulation support