A Service for Statistical Analysis of Marine Data in a Distributed e-Infrastructure

Gianpaolo Coro, Antonio Gioia, Pasquale Pagano, Leonardo Candela
Istituto di Scienza e Tecnologie dell’Informazione A. Faedo - CNR, Pisa, Italy

Context
In the Computational Statistics domain, experts need environments in which they can easily manipulate data and run experiments by themselves. This is the case of Computational Biology, in which data mining procedures are usually applied to big datasets. In such domain, Distributed e-Infrastructures have been yet indicated as possible enabling layers for scientists who want to share data and results with other remote collaborators and run experiments in easy way.

Proposal
We propose a workbench software, named Statistical Manager (SM), that offers a rich array of statistical computations and data mining applications for a variety of biological and marine related problems. Capabilities:
- supplied "as-a-Service" via the D4Science distributed e-Infrastructure;
- parallel and distributed computing of demanding tasks on large datasets;
- comprehensive and open set of off-the-shelf state-of-the-art implementations of approaches to biological problems, that include: environmental features clustering, simulation of biology-related functions, simulation of climatic scenarios, niche modelling;
- integrated with Virtual Research Environments, that support workspace and storage systems to share large datasets, experimental parameters and results;
- exposed for programmatic consumption through Java APIs and SOAP protocol;
- endowed with an automatically generated Web interface.

Features
The Service defines five base operations that can be requested by clients:
- GetCapabilities: to retrieve service metadata (Capabilities), which describe the algorithms supplied by the service.
- GetProcessDescription: to retrieve detailed information about the processes that can be executed through the service. E.g. inputs and outputs descriptions based on an xsd schema.
- Execute: to run one of the processes listed among the Capabilities.
- GetProcessStatus: to check the completion status of a computation.
- GetProcessOutput: to retrieve the process output.

Highlights
SM implements a simple approach to integrate algorithms, including R scripts;
The Web GUI is automatically generated based on the algorithms inputs and outputs declarations. It allows scientists to execute and monitor the algorithms;
The parallel executions through D4Science are transparent to the user;
The features can be mapped onto OGC WPS specifications;
SM is a component of the gCube open source software system, that supports the development of Distributed e-Infrastructures and the definition of Virtual Research Environments.

Contact Person: Gianpaolo Coro, gianpaolo.coro@isti.cnr.it

www.i-marine.eu ~ info@i-marine.eu