

A WEB-BASED PLATFORM FOR EXPERIMENTAL DESIGN AND TIME SERIES ANALYSIS

Fabiano Lever, Federico Scrinzi, Matteo Poletti, Zeni Severnio

The start of this project goes back to the last summer when we took part to a three weeks long “Internet camp” organized by “Fondazione Bruno Kessler”^[1]. There we acquired many skills and remarkable data mining and machine learning knowledge, so when we got home we decided to put it into practice.

Computers have become increasingly fundamental in research. Scientists use them in every step of their activity: to observe a phenomenon, getting information about it, organizing their experiments, to collect and analyze data and to share their results. The application of informatics to fields of research different from the traditional ones - biology for example - helped greatly to the their development.

However the potential that computers have in not always fully exploited. The complexity of modern systems, their costs and the difficulties found in running them have limited it.

We noticed the lack of an easy to use, widely accessible and at the same time highly performing tool that could help scientists to design their experiment and perform analysis on data in the form of time series.

If we could implement a web based platform where experimental design and data mining could be carried out, this hole would be filled to some extent. The advantages of a web based service on a common computer software are many: you don't have to install anything on your computer but you access the service through your web browser, you can reach it from anywhere, the hardware of the server where the software is run can be optimized and you can easily share your data and procedures with other scientists.

To achieve the goal we set ourselves there were two tasks that had to be carried out: implement the data mining and machine learning algorithms we were going to offer and integrate them into a web platform.

The research we did on the potentially useful machine learning tools for data exploration and analysis led us to the decision of implementing a tool clustering and a tool of classification. For the clustering tool we chose and developed two algorithms: the k-Means and K-Medoid, while for the classification task the algorithm that most suited our needs was the k-NN. Both clustering and classification involved distances between time series calculations so we dedicated particular attention to this aspect and we assessed the Dynamic Time Warping (DTW) to be our main topic of research as far as distances were concerned.

In the process of implementation of the algorithms we cared particularly of their performances as they were to be used on very large amounts of data. For

this reason we decided to explore a newest field in scientific computing: the GPGPU. The multi-core architecture of modern GPUs allows to parallelize calculations reducing remarkably the computational time taken by the algorithms.

While still developing the tools described above we started to work at the integration of these into a web based platform. Marco Grimaldi, a researcher of the "Fondazione Bruo Kessler"^[1] indicated to us Galaxy as a possible solution to this problem. Galaxy is a brand new platform developed by the "Penn State University"^[2] to give scientists in the fields of bioinformatics and genetics a starting point for the development of their software and ease down the usage of it for the non experts. The integration of the tools we developed into Galaxy turned out to be pretty straightforward even if we had to rewrite several modules such as the data management and the data display ones to suit our needs. The platform also offers a workflow management tool that is very efficient for experimental design and sharing procedures between scientists.

Finished this we started to load our software on KRK, a workstation that "Fondazione Bruno Kessler"^[1] kindly allowed us to access, and refined our software.

The result of our research is Galaxy-YEPS a revolutionary and remarkably powerful tool for experimental design and time series analysis released under a GNU GPL 3 license. The most significant elements of it are the great power of computation thanks to the GPGPU technology, the high accessibility of the tool as you can use it from anywhere and you don't need to install anything on your computer, it is easy to use even for people with little experience in informatics and it gives researchers the chance to share data and procedures easily thanks to an experimental design tool.

However we think that Galaxy-YEPS is just the "tip of an iceberg" compared to its potentials. We strongly believe that its development could lead to new ways to do science in many fields such as genomics. We already defined many points in which it could be improved and expanded.

The development of Galaxy-YEPS was an extremely positive experience and one last think we want to underline is that throughout the realization of the project, given the distance between the places where we live and that we go to four different schools, we met physically only a few times. The work was discussed and organized through the Internet with a mailing list, a wiki and two repositories where we kept the code we were developing.

A great thank goes to "Fondazione Bruno Kessler"^[1] and to Marco Grimaldi whose advices were of great help for the development of the project.

[1] Fondazione Bruno Kessler is situated in Trentino, a province in northern Italy governed under a special autonomy statute. The foundation, with more than 350 researchers, conducts studies in the areas of Information Technology, Materials and Microsystems, Italo-Germanic studies, and Religious sciences.

[2] "Center for Comparative Genomics and Bioinformatics" Pennsylvania State University.