Research Article

iMedPub Journals
http://www.imedpub.com

Chemical informatics
ISSN 2470-6973

2015
Vol. 1 No. 2: 7

DOI: 10.21767/2470-6973.100007

A Web-based Graphic User Interface of PML for Machine Learning in Parallel Running Head: A Web-based GUI of PML for Machine Learning

Runyu Jing[1], Rong Li[2], Xuemei Pu[1] and Menglong Li[1]

1 College of Chemistry, Sichuan University, Chengdu, China
2 College of Computer Science, Sichuan University, Chengdu, China

Corresponding author: Menglong Li

liml@scu.edu.cn

College of Chemistry, Sichuan University, Chengdu 610064, P. R. China.

Tel: +86-28-89005151
Fax: +86-28-85412356

Citation: Jing R, Rong Li, Xuemei Pu, et al. A Web-based Graphic User Interface of PML for Machine Learning in Parallel Running Head: A Web-based GUI of PML for Machine Learning. Chem Inform. 2015, 1:2.

Abstract

With increasing complexity of the samples, Cheminformatics and Bioinformatics have become powerful tools in assisting experiments. Due to diversity of the data from different fields, researchers usually need to use multiple methods for comparison in order to obtain one optimized model. However, the existing methods rely on different dependent packages and running environments. Therefore, it is time-consuming to integrate the methods together. In order to reduce the time cost of the data modeling and results comparison, we provided PML. Additionally, we developed a web-based graphic user interface by using JavaScript and PHP. By means of the GUI, users can generate the script of PML more easily, and can make certain number of machines in a local area network (LAN) as the computing source for running and controlling PML tasks. We hope that the GUI could simplify the progress of task generation of PML and help researchers improve research efficiency.

Keywords: Machine learning; GUI; Graphic user interface; Distributed computing; Methods comparison; Data mining

Received: August 25, 2015; **Accepted:** September 10, 2015; **Published:** September 14, 2015

Introduction

With the rapid development of analytic technique, machine learning methods are widely used in several fields, such as Cheminformatics and Bioinformatics, for digging useful information from the experimental data. In most cases, the distribution of data from different fields would be variant due to the devices and samples of the experiments. Therefore, to deal with different data, appropriate methods are needed. For example, the data from RNA sequencing would have large number of genes, which would be far more than the number of samples, thus the methods from graphic theory and statistics are necessary to reduce the scale of genes [1-3]. Moreover, if the distribution of the data is not simple, the traditional linear methods usually could not model the data very well [4,5]. In this case, the nonlinear methods, such as kernel trick and sparse factor, are necessary for improving the performance of modeling and predicting [6,7]. Many well developed machine learning tools have been released; however, the tools are hard to integrate due to the distinct running environments. For instance, if we want to use several methods to model a dataset and use cross validation or leave-one-out for the training dataset, the rework of dataset slicing is usually inevitable. Therefore, a tool which could integrate and compare several methods from different environment is necessary.

Based on the motivation, we developed PML, a software which could integrate the methods and running in parallel [8]. Further, to make the server version of PML more user-friendly, we developed the web-based GUI. With this GUI, users could generate and control PML tasks more effectively. In addition, users could set one or more computer as a computing source in a local area network (LAN). The code could be downloaded at http://cic.scu.edu.cn/pml.

The Construct of the GUI

PML for task processing

We developed PML for processing machine learning tasks in parallel. PML can process dimension reduction, grid search, cross validation and result analysis in parallel. Moreover, more than a

single machine, PML could use multiple machines as a cluster to process the tasks by combining BOINC. The output is in HTML format and could be view in browser. Moreover, the results are put in multiple independent folders so that users can move it easily. The mechanisms of fault-tolerant and interrupt recovery are achieved to confirm the stability of the execution of PML. The methods of WEKA and Waffles have been combined into PML, and users could combine new methods into PML through the provided command API. The intermediate data and scripts are archived for repeat, exam or other use. Through PML, users and researchers could modeling data and find the best model more effectively.

The web-based GUI

The input of PML is a script and is submitted by command line. Therefore, the operation would be complicated if users submit a task to a server by SSH. Considering that most of the situations that using PML for large amount of calculation are on a server with several CPU cores or on a cluster with multiple machines, we developed the GUI to simplify the operation of task submission and controlling.

With this GUI, users could create a PML task including data file uploading, method selection, grid search setting and process controlling. Considering that when choosing a method for the first time, users would want to know the details of the method and the related parameters, we provided a floating window to show the brief explanations of the methods, parameters and options. When the task submitted, users could control the process of a task, including stop, continue and delete **(Figure 1)**. After the complement of a task, a link would be provided to view the results. Besides, users could generate a script without submission, and the script could be copied to anywhere.

PML provided the mechanism of grid search, but the input format is not so easy to write. Therefore, we provided some functions to simplify the setting of grid search. By using the GUI, users could modify and view the changed parameters in real time. Additionally, we also provided a brief explanation of the input format in the floating window **(Figure 2)**.

Since PML has two versions, e.g., server and desktop, users could 1) configure a single machine as the computing server by using PML desktop and the GUI or 2) configure multiple machines as a cluster for computing by using PML server and the GUI. The installation of the two versions is different, but after the installation, the setting and usage of the GUI are same. The GUI is written in JavaScript and PHP, thus the installation of this GUI is only to modify the configuring file of Apache. Moreover, we provided manual and script to simplify the installation of GUI.

Conclusion

With the rapid development of analytic technique, the experiment data became increasing complicated. In order to dig the useful information from the data, multiple statistical analysis and machine learning methods become necessary. To improve the efficient of the using and comparison of the methods, we provided PML. Further, to simplify the operation of submission and controlling of the tasks, we developed the GUI. The GUI simplifies the operation of task submission, and provides links to the generated results. We hope that this GUI could save time cost in data modeling and methods comparison, so that researchers could be more efficient in their research.

Acknowledgements

We thank the anonymous reviewers for their patient review and constructive suggestions. This work was supported by the National Natural Science Foundations of China (21375090 and U1230121).

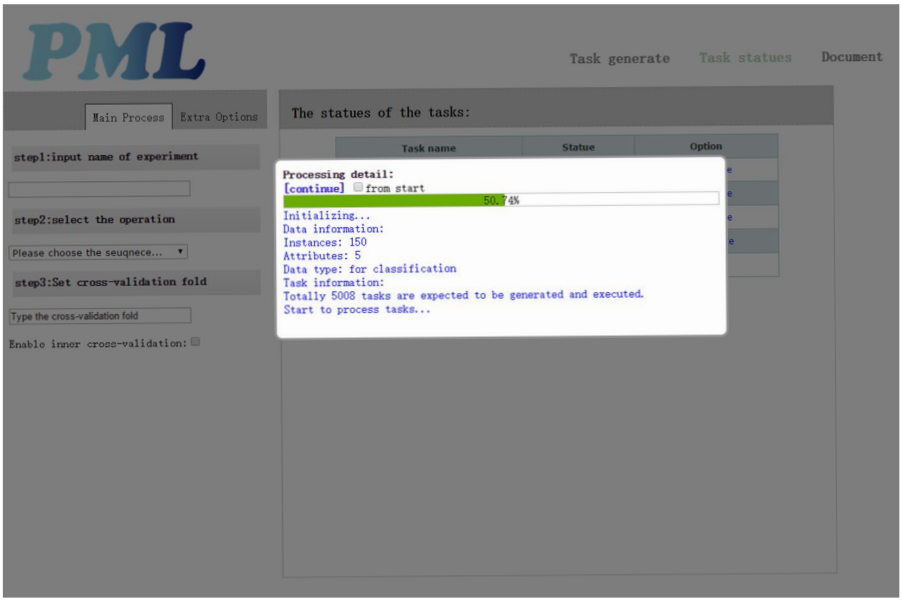

Figure 1 A demonstration of the task process statue control.

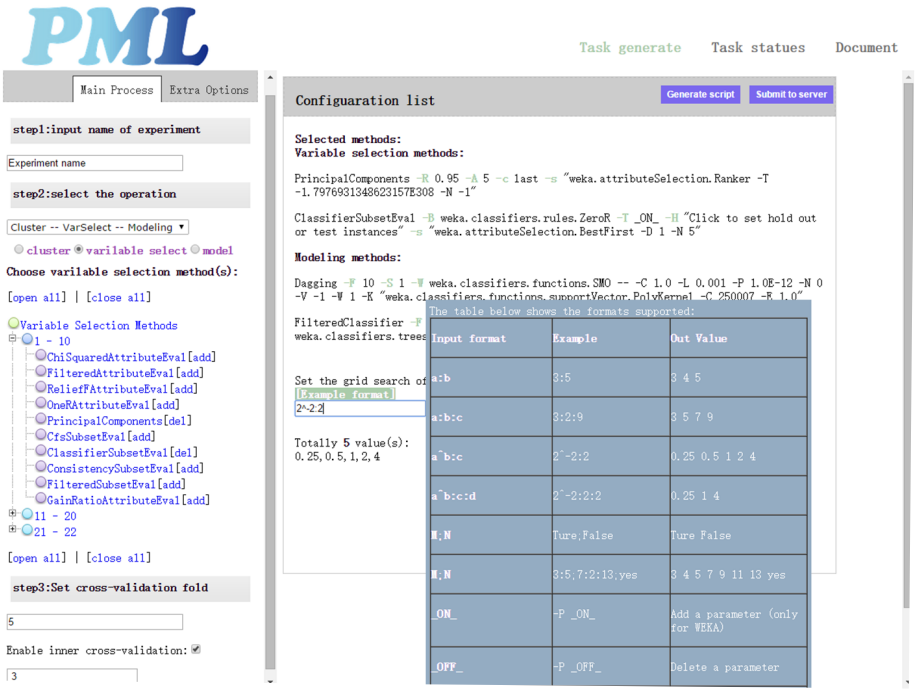

Figure 2 A demonstration of the methods selection and grid search setting.

References

1 Friedman J, Hastie T, Tibshirani R (2008) Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9: 432-441.

2 Lisewski AM, Quiros JP, Ng CL, Adikesavan AK, Miura K, et al. (2014) Supergenomic Network Compression and the Discovery of EXP1 as a Glutathione Transferase Inhibited by Artesunate. Cell 158: 916-928.

3 Su Z, Fang H, Hong H, Shi L, Zhang W, et al. (2014) An investigation of biomarkers derived from legacy microarray data for their utility in the RNA-seq era. Genome Biology 15: 523.

4 Wang Y, Jing R, Hua Y, Fu Y, Dai X, et al. (2014) Classification of multi-family enzymes by multi-label machine learning and sequence-based descriptors. Analytical Methods 6: 6832-6840.

5 Wu Y, Jing R, Lin J, Jiang Y, Kuang Q, et al. (2014) Combination use of protein–protein interaction network topological features improves the predictive scores of deleterious non-synonymous single-nucleotide polymorphisms. Amino Acids 46: 2025-2035.

6 Bhattacharya A, Dunson DB (2011) Sparse Bayesian infinite factor models. Biometrika 98: 291-306.

7 Carvalho CM, West M (2008) High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics. J Am Stat Assoc 103: 1438-1456.

8 Jing R, Sun J, Wang Y, Li M, Pu X (2014) PML: A Parallel Machine Learning Toolbox for Data Classification and Regression. Chemolab 138: 1-6.