Chemometrics and Intelligent Laboratory Systems 138 (2014) 1–6

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Software Description

PML: A parallel machine learning toolbox for data classification and regression

Runyu Jing, Jing Sun, Yuelong Wang, Menglong Li *, Xuemei Pu

College of Chemistry, Sichuan University, Chengdu 610064, PR China

ARTICLE INFO

Article history:
Received 17 April 2014
Received in revised form 2 July 2014
Accepted 9 July 2014
Available online 17 July 2014

Keywords:
Toolbox
Parallel
Cross-platform
Modeling
Data mining
Grid search

ABSTRACT

Motivated by timesaving when dealing with the large-scale calculation for data modeling in parallel with multiple CPU cores or machines and result comparison, PML was designed in this study. PML has the ability to do dimension reduction and grid search in parallel, support both classification and regression, and could generate HTML pages as output for results comparison. Written in PERL, PML is compatible with Windows and Linux. This open source software is free available together with the code, manual, examples and license from http://cic.scu.edu.cn/pml or https://github.com/limlcic/PML.

1. Introduction

In the process of a research in some fields, such as Chemometrics and Bioinformatics, researchers usually need to build some models based on a dataset which is for cross validation, and find out the best fit method(s) for the next independent test to estimate the stability against over-fitting [1–6]. On the other hand, in some studies, a newly developed machine learning method is needed to be compared with others [7–9]. Many published tools, methods and programing languages, such as RapidMiner, KNIME [10], Orange [11], caret [12] and MI-Flex [13], are available for this requirement, but the workflow construction, sub-dataset generation and result comparison are still time-consuming when the number of used modeling methods is not small. In order to decrease the time of data modeling and provide a convenient way for users to use the machine learning methods, we developed PML, a Parallel Machine Learning toolbox. PML mainly focuses on saving time in the process of modeling especially when multiple machine learning methods are used.

PML is available in modeling data with various machine learning methods in parallel, and only needs an input script which could be written by users. Integrating with WEKA [14] and Waffles [15], PML has now gathered more than 20 methods for dimension reduction and more than 80 methods for modeling. Meanwhile, PML also enables the integration of other tools by command line interface. Two sub-versions, Desktop and Server, are provided depending on the diverse requirements. PML-Desktop allows running on one machine in parallel, and PML-Server is designed to support parallel computing among multiple machines. PML-Desktop is compatible with Windows and Linux, and could be installed and executed simply. PML-Server uses the BOINC platform [16], which needs at least one machine as a Server running on Linux with the BOINC-server, and the other machines as Clients running on Windows or Linux. These two sub-versions of PML support the same format of input script when the installation is completed.

PML is under the GNU LESSER GENERAL PUBLIC LICENSE with version 3 and freely available together with code, manual, development document, examples and License from http://cic.scu.edu.cn/pml/ or https://github.com/limlcic/PML.

2. Methods and architecture

The basic workflow of PML is straightforward: Initialization — dimension reduction (optional) — modeling and cross-validation — output generation. Each task which is generated by PML only depends

* Corresponding author. Tel.: +86 28 89005151; fax: +86 28 85412356.
E-mail addresses: jingryedu@gmail.com (R. Jing), sunjinglisa@163.com (J. Sun), wangyuelong2043@gmail.com (Yuelong Wang), liml@scu.edu.cn (M. Li), xmpuscu@scu.edu.cn (X. Pu).

http://dx.doi.org/10.1016/j.chemolab.2014.07.005
0169-7439/

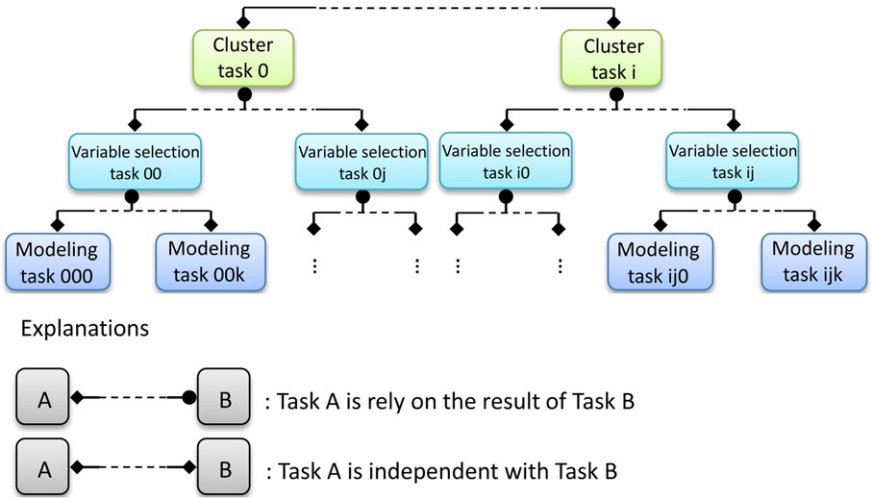

Fig. 1. The relationship of each task in the process of PML execution.

on the outcome of its previous task, thus PML could run in parallel when multiple methods were used (Fig. 1). Besides, the whole workflow of PML is shown in Fig. 2.

PML is cross-platform compatible, and uses only few other PERL modules. Besides, to confirm the stability, PML has been tested on several platforms (detailed in Section 6).

2.1. Parallel

PML-Desktop uses the threads module of PERL to achieve parallel computing, and it could use all the CPU cores of a single machine. PML-Server uses the BOINC platform, the tasks of dimension reduction and modeling could be executed in parallel among the Client machines. However, the tasks of the output analyses can only be executed in parallel on the Server machine.

2.2. Dimensionality reduction

PML supports the data cluster (reduce instances) and variable selection (reduce attributes). The two functions are independent from each other, thus users can decide the sequence of the utilization of the two functions by modifying the input script. Besides, dimensionality reduction tasks would be executed before cross validation, which means that the datasets for cross validation would have the same dimension of attributes.

2.3. Grid search

Most of the methods that are employed by PML have some parameters which could be modified by users. PML allows users to specify one or more parameters and gives them several values by modifying the input script. Then, the comparisons of the outputs from the different

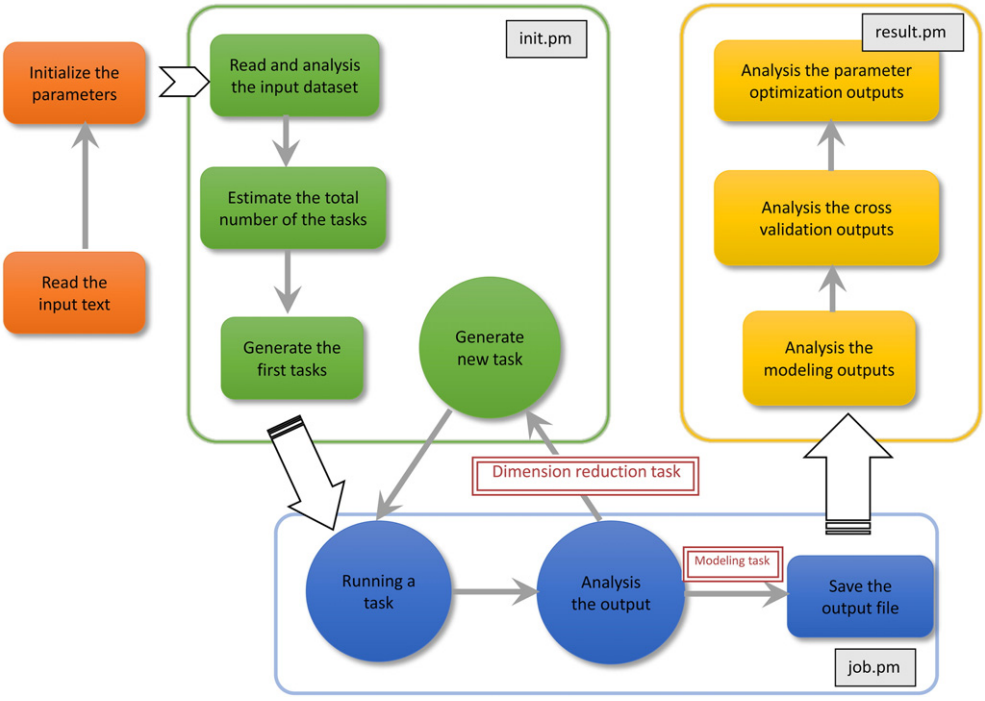

Fig. 2. The workflow and the modules of PML.