PivotalR: A Package for Machine Learning on Big Data

Hai Qian[1,*]

1. Pivotal Inc.
*Contact author: hqian@gopivotal.com

Keywords: big data, machine learning, database, usability

PivotalR [1] is an R package that provides a front-end to PostgreSQL [2] and all PostgreSQL-like databases such as Pivotal Inc.'s Greenplum Database (GPDB) [3], HAWQ [4] on Hadoop. PivotalR also provides the R wrapper for MADlib [5]. MADlib is an open-source library for scalable in-database analytics. It provides data-parallel implementations of mathematical, statistical and machine-learning algorithms for structured and unstructured data. Thus PivotalR also enables the user to apply machine learning algorithms onto big data.

In recent years, Big Data has become an important research topic and a very realistic problem in industry. The amount of data that we need to process is exploding, and the ability of analyzing big data has become the key factor in competition. Big data sets do not fit into computer's memory and it would be really slow if the big data sets were processed sequentially. On the other hand, most contributed packages of R are still strictly sequential, single machine, and they are restricted to small data sets that can be loaded into memory. As computing shifts irreversibly to parallel architectures and big data, there is a risk for the R community to become irrelevant.

PivotalR, which provides an R front-end with data.frame oriented API for R users to access big data stored in distributive databases or Hadoop distributive file system (HDFS). PivotalR puts more emphasis on machine learning by providing a wrapper for MADlib, which is an open-source library of scalable in-database machine learning algorithms. Actually PivotalR offers more than what MADlib has. It adds functionalities that do not exist in MADlib, for example, the support for categorical variables.

PivotalR makes it easier to work on big data sets in databases or HDFS. Many queries that are difficult to construct in SQL client can be easily constructed using PivotalR. This make sit suitable for data preprocessing. PivotalR also makes it easy to create many algorithm prototypes that can directly run in database using familiar R syntax. Besides, PivotalR is portable onto many different platforms, and the prototype code is the same on all supported platforms.

Although PivotalR is targeted at big data, database, and Hadoop, no prior knowledge is needed. The objective of PivotalR is to give the normal R users an easy access to all of these without learning extra knowledge.

References

[1] PivotalR, http://cran.r-project.org/web/packages/PivotalR/index.html. version 0.1.15.1.
[2] PostgreSQL, http://www.postgresql.org/
[3] Greenplum Database, 2013a. http://www.gopivotal.com/products/pivotal-greenplum-database. version 4.2.4.
[4] HAWQ, http://www.gopivotal.com/pivotal-products/pivotal-data-fabric/pivotal-hd. version 1.2 (to be released).
[5] MADlib, http://madlib.net. version 1.5 (to be released).