The data.sample package: sampling as a big data mining tool

M. CORNEC [a] and J. DAS NIEVES [a]

[a] Data Operation Unit
CDISCOUNT
120-126 quai Bacalan – CS 11584
matthieu.cornec@cdiscount.com
jean.dasnieves@discount.com

Mots clefs: big data inference, sampling.

Not a single day without a newspaper article on the big data deluge. According to media coverage, hundreds of Teraoctets (To) would be waiting to be explored, and to deliver great value for consumers and companies. At cdiscount, French leading ecommerce website, we collect a dozen of To on a monthly basis.

At the same time, R, maybe the most popular statistical language is not ready for the Big Data Era. This native drawback is well known since R must load data sets into RAM.

Different strategies have been developed to tackle this challenge. Among them, we can quote: muscling in-house hardware, combining R and a big data relational data base language (such as Hive), the biglm package, the RSQLite package, ff package. Their main purpose is to run SQL like queries for data sets that do not fit into memory. Thus, they give accurate and deterministic results such as the sum or the mean of a variable.

In this poster, we defend the following strategy: as far as data analysis is concerned, sampling is a reliable, fast, and cheap data mining tool for big data. This statement can sound paradoxical since sampling is traditionally associated to the 20th century and to the theory of representative sampling. Nowadays, it is a common belief that we have access to exhaustive piece of information, so why sampling? The reason is the following: when it deals with modeling by opposition to reporting, the error induced is negligible in comparison with model errors, model noise, estimation error,

We introduce the data.sample package whose main function read.table.ds takes the location of big file as input and returns an object of class table.ds, which contains the sampled dataset together with the sampling weights. The main interest is that this strategy is not limited by the RAM size. By allowing the reuse of other R packages, it produces fast and actionable results.

To support our point view, we give theoretical insights following [1,2], simulation studies, and real data sets manipulations.

The data.sample package will be available on request from the authors.

Références

[1] A. Kleiner, A. Talwalkar, P. Sarkar, and M.I. Jordan (2011). Bootstrapping big data. Big Learn, 2011.
[2] Kleiner, A., Talwalkar, A., Sarkar, P., & Jordan, M. I. (2011). A scalable bootstrap for massive data. *arXiv preprint arXiv:1112.5016*.
[3] T. J. Hastie, R.J. Tibshirani, and J.H. Friedman. The elements of statistical learning. Springer, 2009.