Abstract

One of the important limitations of the standard \textit{glm} procedure for estimation of the parameters of logistic regression model in R is the size of the data that is kept in the memory, i.e. the original sample and algorithm-specific temporary data, effectively restricting both cardinality and dimensionality of the sample.

The \textit{biglm} package makes possible to overcome the restriction on the number of observations present in the sample. But it still leaves the dimensionality limitation, due to the second-order algorithm being used to fit the models.

The possibility of use of the first-order, stochastic gradient-descent optimization method creates a tradeoff between the rate of convergence and the maximal dimensionality of the sample. Interestingly, in case of smoothly regularized logistic regression (e.g. L2-based, ridge estimate), it is possible to parallelize the first-order method w.r.t. data sample, greatly improving the computation time of a single iteration and – in practice – reducing the advantage of second-order method.

The gradient-based approach outlined above has been implemented in the Netezza Analytics package using Netezza Performance Server as a database
backend.

In our presentation we will show the performance and accuracy study of the logistic regression implementation provided in the package.