THE UNIVERSITY OF CHICAGO

Department of Statistics

MASTER'S THESIS PRESENTATION

KAILIN LIU

Department of Statistics
The University of Chicago

A Study on the Computing Performance of BLB (Bag of Little Bootstraps) on Large Datasets

THURSDAY, February 21, 2013, at 10:00 AM

110 Eckhart Hall, 5734 S. University Avenue

ABSTRACT

Assessing the quality of estimators is a task of fundamental importance in Data Science. Quality assessments such as confidence region, bias, risk and so on provide much more information than a simple point estimate. This paper aims to study a statistical sampling algorithm Bag of Little Bootstraps (BLB) that solves the same class of problems as general bootstrapping, but which parallelizes better. We do this by working on a Scala implementation that runs on the Spark cluster computing framework as described in "A Scalable Bootstrap for Massive Data," as well as a Python expression which can sample gigabyte datasets with performance comparable to hand-tuned parallel code. We also tried to evaluate the computing performance of this algorithm by performing model verification of a SVM classifier on a subset of the dataset Enron email corpus.
