NeuroData SIMPLEX Report: February 2017

The following report documents the progress made by the labs of PI Joshua T. Vogelstein and Co-PIs Randal Burns and Carey Priebe at Johns Hopkins University towards goals set by the DARPA SIMPLEX grant.

Contents

Data: What's in the Cloud

Statistical Theory and Methods

LOL @jovo

LOL took a backseat this month while jovo had a baby :)

Multiscale Generalized Correlation (MGC)

Your content here. Please make sure that it fits on one page.

Figure 1: Please provide a detailed caption for your figure.

RerF

Your content here. Please make sure that it fits on one page.

Figure 2: Please provide a detailed caption for your figure.

Discriminability

Your content here. Please make sure that it fits on one page.

Figure 3: Please provide a detailed caption for your figure.

Low-rank Assumption Discussion

Robust Law of Large Graphs

Your content here. Please make sure that it fits on one page.

Figure 4: Please provide a detailed caption for your figure.

Nonparametric Network Dependence Test

Your content here. Please make sure that it fits on one page.

Figure 5: Please provide a detailed caption for your figure.

Batch effect removal in dimension reduction of multiway array data

Your content here. Please make sure that it fits on one page.

Figure 6: Please provide a detailed caption for your figure.

Reduced Dimension Clustering

Your content here. Please make sure that it fits on one page.

Figure 7: Please provide a detailed caption for your figure.

Table 1: The runtime and memory consumption of FlashR on the billion-scale datasets on the 48 CPU core machine. The runtime of iterative algorithms is measured when the algorithms converge. We run PageRank on the PageGraph dataset, run k-means on PageGraph-32ev and the remaining algorithms on Criteo.

	Runtime (s)	Memory (GB)
Correlation	91.23	1.5
PCA	136.71	1.5
NaiveBayes	76.55	3
LDA	2280	8
Logistic regression	4154.40	26
k-means	1110.82	28
PageRank	3900	135

Scalable Algorithm Implementations

FlashX

We use FlashR to process the billion-scale datasets to demonstrate its scalability (Table 1). We use three datasets here: (i) the Criteo dataset has over four billion data points with binary labels (click vs. no-click), used for advertisement click prediction; (ii) PageGraph is the adjacency matrix of a graph, which has 3.5 billion vertices and 128 billion edges; (iii) PageGraph-32ev are 32 singular vectors that we computed on the largest connected component of Pagegraph with the tools we built previously. In these experiments, we run the iterative algorithms (Logistic regression, k-means and PageRank) on the datasets until they converge.

Even though we process the billion-scale datasets in a single machine (with 48 CPU cores), none of the algorithms are prohibitively expensive. Simple algorithms, such as Naive Bayes and PCA, require one or two passes over the datasets and take only one or two minutes to complete. Logistic regression and k-means take about $10-20$ iterations to converge. Because the PageRank implementation uses the power method, it takes 100 iterations to converge. Nevertheless, all of the iterative algorithms take about one hour or less.

FlashR scales to datasets with billions of data points easily when running out of core. Most of the algorithms have negligible memory consumption. PageRank consumes more memory because the sparse matrix multiplication in PageRank keeps vectors in memory for semi-external memory computation. The scalability of FlashR is mainly bound by the capacity of SSDs. The functional programming interface generates a new matrix in each matrix operation, which potentially leads to high memory consumption. Thanks to lazy evaluation and virtual matrices, FlashR only needs to materialize the small matrices to effectively reduce memory consumption.

ndstore

Your content here. Please make sure that it fits on one page.

Figure 8: Please provide a detailed caption for your figure.

ndviz

Your content here. Please make sure that it fits on one page.

Figure 9: Please provide a detailed caption for your figure.

knor

Your content here. Please make sure that it fits on one page.

Figure 10: Please provide a detailed caption for your figure.

Scientific Pipelines: Infrastructure & Dataset Specific Progress

SIC

ndstore

ndmg

ndviz

MRI

CLARITY

Ophys

Bibliography

Manuscripts

[1] N. Peeps, “Paper with a cool title,” 2017.

Invited Talks

[1] N. Peeps, “Talk with a cool title,” SIAM, JSM, NIPS, ???, Jan 2017.

Conferences

[1] N. Peeps, “Poster with a cool title,” SIAM, JSM, NIPS, ???, Jan 2017.

References