Nonparametric Distributed Learning Framework: Algorithm and Application to Variable Selection

Scott Bruce, Zeda Li, Hsiang-Chieh (Alex) Yang, Subhadeep Mukhopadhyay

Temple University - Student Award Seminar

April 24, 2015
Motivating Example - Personalized Expedia Hotel Searches

**Goal:** Expedia is interested in identifying important variables that could lead to efficient personalized hotel rankings for individual online users, thus maximizing purchases for given hotel queries.

**Key Challenges**

- **Volume:** Over 10 million observations across 52 variables.
- **Variety:** Different data types require different measures of significance.
- **Scalability:** Distributed, parallel processing for massive data analysis.
Statistical Challenge I: Volume

**Big Data 1.0 - Wide Data**

- **Tall data**: Hundreds of variables but millions/billions of samples.
- Data cannot be processed *all at once* using standard statistical software.
- Currently NO statistical approach available to address this problem.

**Big Data 2.0 - Tall Data**
Introduction

Statistical Challenges

MetaLP Algorithm

Statistical Challenge II: Variety

Unsolved problems of Data Science:

- Traditional statistical modeling develops tools that are specific for each data type.
- Implementation gets dauntingly complex for large heterogeneous datasets as they require the data-type information for each variable to compute the proper statistic.
- Goal: Automation for Mixed-data based on united computing formula.

<table>
<thead>
<tr>
<th>Independent variable</th>
<th>Categorical – binary</th>
<th>Categorical – nominal</th>
<th>Continuous – normally distributed</th>
<th>Continuous – non-normally distributed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categorical – binary</td>
<td>Fisher exact/chi-square</td>
<td>Chi square</td>
<td>Independent samples t-test</td>
<td>Mann-Whitney test</td>
</tr>
<tr>
<td>Categorical – nominal</td>
<td>Chi square</td>
<td>Chi square</td>
<td>ANOVA</td>
<td>Kruskal-Wallis test</td>
</tr>
<tr>
<td>Continuous – normally distributed</td>
<td>Independent samples t-test</td>
<td>ANOVA</td>
<td>Pearson’s correlation/linear regression</td>
<td>Spearman’s correlation</td>
</tr>
<tr>
<td>Continuous – non-normally distributed</td>
<td>Mann-Whitney test</td>
<td>Kruskal-Wallis test</td>
<td>Spearman’s correlation</td>
<td>Spearman’s correlation</td>
</tr>
</tbody>
</table>

†, Fisher exact test is used when one or more of the expected counts in a 2 x 2 table is small i.e. less than five.

Statistical Challenge III: Scalability

- Scalability is crucial to efficiently analyze datasets like the Expedia hotel search dataset.
- Distributed, parallel processing capabilities (e.g. MapReduce) greatly improve the speed of data analysis.
- Can we construct a statistical inference framework that can leverage the computational efficiency of the MapReduce functionality?
LP Nonparametric Parallelizable Algorithm: MetaLP

1. **Partition**
   Assign observations to different subpopulations in a reasonable manner (e.g., random assignment). This step can be omitted if the dataset already follows a predefined partitioning scheme.

2. **Map**
   Estimate the LP statistic and corresponding Confidence Distribution for each subpopulation. LP Hilbert functional inner-product representation:

\[
LP_{\ell}[j; X, Y] = \text{Cor}[T_j(X; X), Y] \\
= \mathbb{E}[T_j(X; X) T_1(Y; Y)] \quad \ell = 1, \ldots, k. \quad (1)
\]

\[
H(LP_{\ell}[j; X, Y]) \\
= \Phi \left[ \sqrt{n} \left( LP_{\ell}[j; X, Y] - \widehat{LP}_{\ell}[j; X, Y] \right) \right] \quad (2)
\]
**LP Nonparametric Parallelizable Algorithm: MetaLP**

### Regularization and Combine

Compute the combined meta-analysis based regularized confidence-distribution for LP Statistics that accounts for the heterogeneity across subpopulations.

\[
H^{(c)}(LP_\ell[j; X, Y]) = \Phi \left[ \left( \sum_{\ell=1}^{k} \frac{1}{\tau^2 + (1/n_\ell)} \right)^{1/2} (LP_\ell[j; X, Y] - \widehat{LP}^{(c)}_\ell[j; X, Y]) \right],
\]

where,

\[
\widehat{LP}^{(c)}_\ell[j; X, Y] = \frac{\sum_{\ell=1}^{k} (\tau^2 + (1/n_\ell))^{-1} LP_\ell[j; X, Y]}{\sum_{\ell=1}^{k} (\tau^2 + (1/n_\ell))^{-1}}.
\]
Flowchart of MetaLP Big-Data Analysis Scheme

Output
Point and interval LP-effect size estimates for each variable.
Four Fundamental Ideas


