Distributed Computing for Privacy-Preserving Data Analysis

Shuang Wang PhD
Overview

- Introduction
- Data Sharing Models
- Privacy-preserving Distributed Computing
- Privacy-preserving Shared Access
- Conclusions
Introduction

• Biomedical science is moving towards data-driven approaches
  » EHR data is being increasingly used
    • pSCANNER: integrating data over 21 million patients
  » Genome-wide association studies (GWAS)
    • Kawasaki Disease
  » Next-Gen Sequencing (NGS) data
    • Illumina’s HiSeq X Ten provides the first $1000 genome sequencing at 30x coverage
• GWAS studies of Kawasaki disease

  » A robust detection of novel loci that is likely to be associated with the Kawasaki disease may require thousands of samples across institutions in different countries (e.g., US, UK and Singapore)

  » But this requires combining individual-level data from multiple sites, which is against institutional privacy policy.
Data Sharing Models

Overview

- Introduction
- Data Sharing Models
- Privacy-preserving Distributed Computing
- Privacy-preserving Shared Access
- Conclusions
Distributed Data Analysis

Dataset A

Dataset B
Distributed Data Analysis

Dataset A
Intermediary statistics

Dataset B
Intermediary statistics
Distributed Data Analysis
Distributed Data Analysis

Dataset A
Intermediary statistics

Global Parameters

Dataset B
Intermediary statistics
Distributed Data Analysis

Dataset A

Intermediary statistics

Global Parameters

Dataset B

Intermediary statistics
Distributed Data Analysis

Loop n times: until global parameters converge

Dataset A

Intermediary statistics

GLORE

Dataset B

Intermediary statistics

Global Parameters

Dataset A

Intermediary statistics

GLORE

Dataset B

Intermediary statistics

Global Parameters
• Suppose $m-1$ features are consistent over $k$ sites

• In each iteration, intermediary results of a $mxm$ matrix and a $m$-dimensional vector are transmitted to $k-1$ sites

Online learning capability to avoid the need for training on the entire database when a single record is updated.

Model learning from distributed sources without sharing raw data.

Institution 2

Patient data

Institution 3

Patient data

Institution 1

Patient data

Strong privacy protection.

Client sites could dynamically shift from online to offline.

WebGLORE: Web-based Grid binary LOgistic REgression

Web-based Grid binary LOgistic REgression (WebGLORE)

Navigation

- **Login**
  - Log into the GLORE system

- **Home**
  - View your GLORE profile page

- **Instructions**
  - Learn the fundamentals of using GLORE

- **Registration**
  - Register an account in GLORE

- **Create Task**
  - Create a new GLORE task

- **WaitForParticipants**
  - Wait for other participants

- **Computation**
  - Computation process

- **Team**
  - Team members

Instructions & Example data

- **PLAYLIST**
  - WebGLORE: build a global predictive logistic regression model

- **Introduction**

Other materials:

- The user manual is available [here](#).
- The Workflow diagram is available [here](#).

Example data for training and testing (save to a local directory):

- IPDLR_1, IPDLR_2
- IPDLR_3, IPDLR_4

55 countries, 800+ visits

WebDISCO: Web-based DIStributed COx Regression Model

WebDISCO is a webservice for biomedical researchers to build a global predictive cox regression model without sharing data. The tool leverages a distributed Newton-Raphson algorithm and an easy-to-use interface to exchange aggregated statistics from participating institutions, which are less privacy sensitive compared to the raw data, to overcome the regulation barriers. The results are guaranteed to be accurate as if models are trained from combined raw data in a central repository. WebDISCO is the first-of-its-kind that enables iterative optimization procedures to be executed over the network in realtime. Meaningful use of WebDISCO can improve statistical power, speedup discovery, and make a difference to applications where sample size matters.

COX Java backbone:
- Check out the source code from here using subversion. (README)

COX Web Service:
- WebDISCO is an extension of the Cox Java backbone
- Check out the source code from here using subversion. (DEPLOYMENT(Same as WebGlrire))

• Extend and scale up our web based secure distributed computing
  » Leverage Linux Virtual machines (VM) in a cloud environment
  » Develop distributed algorithms tailored for GWAS

• To enable ready-to-run analysis, we are creating and testing
  » Pre-built VM images in iDASH cloud environment
  » Automatic provisioning and configuration scripts in VMs
  » Secure computing among multiple VMs.
We will launch these in our collaborating sites (Atlanta, Singapore and London) next year for GWAS in Kawasaki Disease.

- Download pre-built VM images at each site
- Configure network (e.g. IP address)
- Attach local data into the VM
- Consent to start secure computation
Overview

- Introduction
- Data Sharing Models
- Privacy-preserving Distributed Computing
- Privacy-preserving Shared Access
- Conclusions
Remote data analysis service hosted in a secure Cloud

Secured and restricted data set

<table>
<thead>
<tr>
<th>ID</th>
<th>Sex</th>
<th>Age</th>
<th>Blood pressure</th>
<th>Heart disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>M</td>
<td>83</td>
<td>167/96</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>F</td>
<td>32</td>
<td>150/87</td>
<td>Yes</td>
</tr>
<tr>
<td>3</td>
<td>F</td>
<td>37</td>
<td>120/79</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>M</td>
<td>63</td>
<td>105/64</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>F</td>
<td>18</td>
<td>147/90</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Number of women with heart disease? Count Result: 12
WITNESS: A Web Interface for Interactive Survival Studies

Available at https://witness.ucsd-dbmi.org/
WITNESS: A Web InTerface for iNtEractive Survival Studies

Call:
coxph(formula = Surv(cvdttime, censored) ~ ., data = dat, singular.ok = TRUE, method = "breslow")

n = 200, number of events = 140

| Variable | coef  | exp(coef) | se(coef) | z     | Pr(>|z|) |
|----------|-------|-----------|----------|-------|----------|
| age      | -0.00780 | 0.99223  | 0.01196  | -0.65 | 0.51     |
| sbp      | 0.00121  | 1.00121  | 0.00746  | 0.16  | 0.87     |
| dbp      | 0.00323  | 1.00323  | 0.01023  | 0.32  | 0.75     |
| USHTNYes | 0.11598  | 1.12297  | 0.42144  | 0.28  | 0.78     |
| sexMale  | -0.06064 | 0.94117  | 0.17771  | -0.34 | 0.73     |
| smokeYes | -0.16179 | 0.85062  | 0.19149  | -0.84 | 0.40     |
| htnrxYes | -0.06360 | 0.93838  | 0.38039  | -0.17 | 0.87     |

exp(coef) exp(-coef) lower .95 upper .95

<table>
<thead>
<tr>
<th>Variable</th>
<th>exp(coef)</th>
<th>exp(-coef)</th>
<th>lower .95</th>
<th>upper .95</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>0.992</td>
<td>1.008</td>
<td>0.969</td>
<td>1.02</td>
</tr>
<tr>
<td>sbp</td>
<td>1.001</td>
<td>0.999</td>
<td>0.987</td>
<td>1.02</td>
</tr>
<tr>
<td>dbp</td>
<td>1.003</td>
<td>0.997</td>
<td>0.983</td>
<td>1.02</td>
</tr>
<tr>
<td>USHTNYes</td>
<td>1.123</td>
<td>0.890</td>
<td>0.492</td>
<td>2.57</td>
</tr>
<tr>
<td>sexMale</td>
<td>0.941</td>
<td>1.063</td>
<td>0.664</td>
<td>1.33</td>
</tr>
<tr>
<td>smokeYes</td>
<td>0.851</td>
<td>1.176</td>
<td>0.584</td>
<td>1.24</td>
</tr>
<tr>
<td>htnrxYes</td>
<td>0.938</td>
<td>1.066</td>
<td>0.445</td>
<td>1.98</td>
</tr>
</tbody>
</table>

Concordance= 0.568  (se = 0.045 )
Rsquare= 0.009  (max possible= 0.998 )
Likelihood ratio test= 1.82  on 7 df,  p=0.969
Wald test   = 1.81  on 7 df,  p=0.97
Score (logrank) test = 1.81  on 7 df,  p=0.97

Supported by the NIH Grant U54 HL108460 to the University of California, San Diego
Conclusions

• Privacy-preserving Distributed Computing
  » Enable privacy-preserving collaboration across multiple Institutions
  » Exchange less-sensitive aggregated information instead of sensitive patient level data
  » Improve web-based computing with VM-based distributed cloud computing

• Privacy-preserving Shared Access
  » Data never leave the private cloud
  » Approved researchers can perform approved analysis remotely
  » Approved researchers can get only approved results
This study is supported by grants from NIH.

Members in the UCSD team: Lucila Ohno-Machado, Claudiu Farcas, Xiaoqian Jiang, Jihoon Kim, Yuan Wu, Lichang Wang, Li Xiong, Wei Wei, Zhanglong Ji, Challen Lu, Pinghao Li, Wenchao Jiang, Jane Burns.

Members in the OUHSC team: Elisa T. Lee, Wenyu Wang, Fawn Yeh, and Ying Zhang.
Thank you!

Questions?