Big Brain Data Science and Predictive Health Analytics

Ivo D. Dinov
Statistics Online Computational Resource
Health Behavior & Biological Sciences
Computational Medicine & Bioinformatics
Michigan Institute for Data Science
University of Michigan
www.SOCR.umich.edu

Outline

- Driving biomedical & health challenges
- Common characteristics of Big Brain Data
- Data science & predictive analytics
- Case-studies
  - Applications to Neurodegenerative Disease
  - Data Dashboarding
- Compressive Big Data Analytics (CBDA)
Driving Biomedical/Health Challenges

- **Neurodegeneration:** Structural Neuroimaging in Alzheimer’s Disease illustrates the Big Data challenges in modeling complex neuroscientific data. 808 ADNI subjects, 3 groups: 200 subjects with Alzheimer’s disease (AD), 383 subjects with mild cognitive impairment (MCI), and 225 asymptomatic normal controls (NC). The 80 neuroimaging biomarkers and 80 highly-associated SNPs.

![Neurodegeneration Image]

http://DSPA.predictive.space
Moon, Dinov, et al. (2015)

Driving biomedical/health challenges

- **Phenotype-Genotype-Environmental**
  - **Phenotype:** Outward manifestations of the form, shape, or characteristics of specific individuals, cohorts, or morbid conditions
  - **Nature/Genome:**
    - Read the genome → protein synthesis → 2 basic functions → structural proteins determine physical traits or functional traits of protein enzymes catalyzing chemical reactions
  - **Nurture/Environment:**
    - Ambient environment directly effects Pheno + Geno → Provides raw materials needed for the synthetic processes controlled by Nature/Genes → Organisms either synthesize or obtain amino acids with their diet

![Phenotype-Genotype-Environmental Diagram]
Characteristics of Big Biomed Data

IBM Big Data 4V’s: Volume, Variety, Velocity & Veracity

<table>
<thead>
<tr>
<th>Big Bio Data Dimensions</th>
<th>Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>Harvesting and management of vast amounts of data</td>
</tr>
<tr>
<td>Complexity</td>
<td>Wranglers for dealing with heterogeneous data</td>
</tr>
<tr>
<td>Incongruency</td>
<td>Tools for data harmonization and aggregation</td>
</tr>
<tr>
<td>Multi-source</td>
<td>Transfer and joint modeling of disparate elements</td>
</tr>
<tr>
<td>Multi-scale</td>
<td>Macro to meso to micro scale observations</td>
</tr>
<tr>
<td>Incomplete</td>
<td>Reliable management of missing data</td>
</tr>
</tbody>
</table>

Example: analyzing observational data of 1,000’s Parkinson’s disease patients based on 10,000’s signature biomarkers derived from multi-source imaging, genetics, clinical, physiologic, phenomics and demographic data elements

Software developments, student training, service platforms and methodological advances associated with the Big Data Discovery Science all present existing opportunities for learners, educators, researchers, practitioners and policy makers

Dinov, et al. (2016) PMID:26918190
Data science & predictive analytics

- **Data science**: an emerging extremely transdisciplinary field - bridging between the theoretical, computational, experimental, and biosocial areas. Deals with enormous amounts of complex, incongruent and dynamic data from multiple sources. Aims to develop algorithms, methods, tools and services capable of ingesting such datasets and supplying semi-automated decision support systems.

- **Predictive analytics**: process utilizing advanced mathematical formulations, powerful statistical computing algorithms, efficient software tools and web-services to represent, interrogate and interpret complex data. Aims to forecast trends, cluster patterns in the data, or prognosticate the process behavior either within the range or outside the range of the observed data (e.g., in the future, or at locations where data may not be available).

---

Case-Studies – ALS

- Identify predictive classifiers to detect, track and prognosticate the progression of ALS (in terms of clinical outcomes like ALSFRS and muscle function)
- Provide a decision tree prediction of adverse events based on subject phenotype and 0-3 month clinical assessment changes

### Data Source

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Sample Size/Data Type</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>ProAct Archive</td>
<td>Over 100 variables are recorded for all subjects including: Demographics: age, race, medical history, sex; Clinical data: Amyotrophic Lateral Sclerosis Functional Rating Scale (ALSFRS), adverse events, onset_delta, onset_site, drugs use (riluzole)</td>
<td>The PRO-ACT training dataset contains clinical and lab test information of 8,635 patients. Information of 2,424 study subjects with valid gold standard ALSFRS slopes used for processing, modeling and analysis</td>
</tr>
</tbody>
</table>

The time points for all longitudinally varying data elements are aggregated into signature vectors. This facilitates the modeling and prediction of ALSFRS slope changes over the first three months (baseline to month 3).

### Methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Linear Regression</th>
<th>Random Forest</th>
<th>BART</th>
<th>SuperLearner</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>0.081</td>
<td>0.174</td>
<td>0.225</td>
<td>0.178</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.619</td>
<td>0.587</td>
<td>0.568</td>
<td>0.585</td>
</tr>
<tr>
<td>Correlation</td>
<td>0.298</td>
<td>0.434</td>
<td>0.485</td>
<td>0.447</td>
</tr>
</tbody>
</table>
Case-Studies – Parkinson’s Disease

- Predict the clinical diagnosis of patients using all available data (with and without the UPDRS clinical assessment, which is the basis of the clinical diagnosis by a physician)
- Compute derived neuroimaging and genetics biomarkers that can be used to model the disease progression and provide automated clinical decisions support
- Generate decision trees for numeric and categorical responses (representing clinically relevant outcome variables) that can be used to suggest an appropriate course of treatment for specific clinical phenotypes

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Sample Size/Data Type</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPMI Archive</td>
<td>Demographics: age, medical history, sex. Clinical data: physical, verbal learning and language, neurological and olfactory, UPSIT, UPDRS scores, ADL, GDS-15, ... Imaging data: structural MRI. Genetics data: APOE genotypes e2/e3. Cohorts: Group 1 = (PD Subjects), N1 = 263; Group 2 = (PD Subjects with Scans without Evidence of a Dopaminergic Deficit (SWEDD)), N2 = 40; Group 3 = (Control Subjects), N3 = 127.</td>
<td>The longitudinal PPMI dataset including clinical, biological and imaging data (screening, baseline, 12, 24, and 48 month follow-ups) may be used conduct model-based predictions as well as model-free classification and forecasting analyses</td>
</tr>
</tbody>
</table>

Predictive Big Data Analytics: Applications to Parkinson’s Disease

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Positive Predictive Value</th>
<th>Negative Predictive Value</th>
<th>Log Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>AdaBoost</td>
<td>0.996324</td>
<td>0.994141</td>
<td>0.998264</td>
<td>0.9980392</td>
<td>0.9948097</td>
<td>11.4882058</td>
</tr>
<tr>
<td>SVM</td>
<td>0.985294</td>
<td>0.994140</td>
<td>0.977431</td>
<td>0.9750958</td>
<td>0.9946996</td>
<td>8.902166</td>
</tr>
</tbody>
</table>

Dinov, et al., (2016) PMID:27494614
Open-Science & Collaborative Validation

End-to-end Big Data analytic protocol jointly processing complex imaging, genetics, clinical, demo data for assessing PD risk

- Methods for rebalancing of imbalanced cohorts
- ML classification methods generating consistent and powerful phenotypic predictions
- Reproducible protocols for extraction of derived neuroimaging and genomics biomarkers for diagnostic forecasting

https://github.com/SOCR/PBDA

PD Patient Falls – Predictive BD Analytics

- Goals
  1) Harmonize and aggregate complex, multisource, and multi-site PD data
  2) Identify highly predictive features associated with specific clinical traits, e.g., falls
  3) Forecast falls using ML techniques and validate using statistical methods

- Study Design
**PD Patient Falls – Predictive BD Analytics**

**Results**

Missing data pattern

Udall vs. Tel-Aviv data

KS tests on 126 features comparing the distributions. Red horizontal line is the cutoff of $-\log(\alpha)$, where $\alpha$ (desired FDR) = 0.01.

**Udall data differences in MDS_TREM (p = 0.5465), MDS_PIGD (p < 0.001), H and Y scale (p < 0.001), gaitSpeed_Off (p < 0.001) between patients with/without falls**

---

**Feature selection** for the Udall data using RF and KO.

7 common features selected by both methods: MDS_PIGD, gaitSpeed_Off, MOT_EDL, NON_MOTOR_EDL, walk, pos_stab

**Binary classification of falls/no-fall** (5-fold CV) using top 10 selected features (gaitSpeed_Off, ABC, BMI, PIGD_score, X2.11, partII_sum, Attention, DGI, FOG_Q, H_and_Y_OFF)

---

<table>
<thead>
<tr>
<th>Methods</th>
<th>acc</th>
<th>sens</th>
<th>spec</th>
<th>spec</th>
<th>ppv</th>
<th>npv</th>
<th>lor</th>
<th>Interpreter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logit</td>
<td>0.728</td>
<td>0.537</td>
<td>0.855</td>
<td>0.710</td>
<td>0.736</td>
<td>1.920</td>
<td>336</td>
<td>pos_stab</td>
</tr>
<tr>
<td>Random Forests</td>
<td>0.796</td>
<td>0.683</td>
<td>0.871</td>
<td>0.778</td>
<td>0.806</td>
<td>2.677</td>
<td>324</td>
<td>Caudate_DA</td>
</tr>
<tr>
<td>AdaBoost</td>
<td>0.689</td>
<td>0.600</td>
<td>0.742</td>
<td>0.610</td>
<td>0.742</td>
<td>1.902</td>
<td>322</td>
<td>Caudate_DA</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.699</td>
<td>0.707</td>
<td>0.694</td>
<td>0.604</td>
<td>0.782</td>
<td>1.699</td>
<td>320</td>
<td>MOT_EDL</td>
</tr>
<tr>
<td>SVM</td>
<td>0.709</td>
<td>0.561</td>
<td>0.806</td>
<td>0.657</td>
<td>0.735</td>
<td>1.672</td>
<td>318</td>
<td>gait</td>
</tr>
<tr>
<td>Neural Network</td>
<td>0.699</td>
<td>0.610</td>
<td>0.758</td>
<td>0.625</td>
<td>0.746</td>
<td>1.588</td>
<td>318</td>
<td>gender</td>
</tr>
<tr>
<td>Super Learner</td>
<td>0.738</td>
<td>0.683</td>
<td>0.774</td>
<td>0.667</td>
<td>0.787</td>
<td>1.999</td>
<td>318</td>
<td>depression</td>
</tr>
</tbody>
</table>
Case-Studies – General Populations

- UK Biobank – discriminate between HC, single and multiple comorbid conditions
- Predict likelihoods of various developmental or aging disorders
- Forecast cancer

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Sample Size/Data Type</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>UK Biobank</strong></td>
<td></td>
<td>The longitudinal archive of the UK population (NHS)</td>
</tr>
</tbody>
</table>

**Data Source & Sample Size/Data Type**

- **Clinical data:** > 4K features
- **Imaging data:** T1, resting-state fMRI, task fMRI, T2_FLAIR, dMRI, SWI

**Genetics data**

- The longitudinal archive of the UK population (NHS)

**Features**

- **Missing Clinical & Phenotypic data for 10K subjects with sMRI, for which we computed 1,500 derived neuroimaging biomarkers.**
- Including only features observed >30%
  
  **(9,914 × 1,475)**

**Missing Count**

- **Features**

**Cases**

**General Populations**

- **UK Biobank**
  - **Demographics:** > 500K cases
  - **Clinical data:** > 4K features
  - **Imaging data:** T1, resting-state fMRI, task fMRI, T2_FLAIR, dMRI, SWI
  - **Genetics data**

**Case-Studies – UK Biobank (Complexities)**

**Missing values of A2**

**Missing Count**

- **Features**
Case-Studies – UK Biobank – NI Biomarkers

Case-Studies – UK Biobank – Successes/Failures
End-to-end Pipeline Workflow Solutions

SOCR Big Data Dashboard

- Web-service combining and integrating multi-source socioeconomic and medical datasets
- Big data analytic processing
- Interface for exploratory navigation, manipulation and visualization
- Adding/removing of visual queries and interactive exploration of multivariate associations
- Powerful HTML5 technology enabling mobile on-demand computing

Husain, et al., 2015, PMID:26236573
SOCR Dashboard (Exploratory Big Data Analytics): Data Fusion

http://socr.umich.edu/HTML5/Dashboard

SOCR Dashboard (Exploratory Big Data Analytics): Associations
Compressive Big Data Analytics (CBDA)

- Foundation for Compressive Big Data Analytics (CBDA)
  - Iteratively generate random (sub)samples from the Big Data collection
  - Then, using classical techniques to obtain model-based or non-parametric inference based on the sample
  - Next, compute likelihood estimates (e.g., probability values quantifying effects, relations, sizes)
  - Repeat – the process continues iteratively until a criterion is met – the (re)sampling and inference steps many times (with or without using the results of previous iterations as priors for subsequent steps)

Dinov, 2016, PMID: 26998309

CBDA Framework
FAIR Data & Open-Science Principles

- Share resources
- Collaborate
- Permissive licenses (e.g., LGPL/CC-BY)
- Project management (e.g., GitHub/Jira)
- Open-access pubs
- Public-private partnerships
- Co-mentoring of trainees
- Effective transdisciplinary methods
- Resource Interoperability
- Result Reproducibility

Clinical Decision Support

- Hospital Admissions
  - Survival Inference and Clinical outcome forecasting using hospital admissions data (N~60K and k=9):
    - Admission_Len: Duration of hospital stay (in days)
    - Death: Indicator of Death (1) or survival (0)
    - ...
    - Demographics & (Human-labeled) Diagnoses

VarImpPlot: RF Survival Prediction

<table>
<thead>
<tr>
<th>Variable</th>
<th>MeanDecreaseAccuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>AdmissionLengthDays</td>
<td>0.0937 963</td>
</tr>
<tr>
<td>Admission_Type</td>
<td>1.722 151</td>
</tr>
<tr>
<td>Race</td>
<td></td>
</tr>
</tbody>
</table>

Accuracy: 0.8568759
95% CI: (0.85, 0.86)
Sensitivity: 0.9322638
Specificity: 0.1395476
Pos Pred Value: 0.9126514
Neg Pred Value: 0.1729668

NeuralNet Deep Learning: Survival Probability (N=11,773)

Test/Validation Cases
Clinical Decision Support

- Population-wide Study of Health and Disease
- National Health Service (NHS)
- Longitudinal Data (N>1M, k>4,300)

Personalized medicine – Traumatic Brain Injury (TBI)

LONI/USC, BIRC/UCLA, SOCR/UMich
Acknowledgments

Funding
NIH: P20 NR015331, U54 EB020406, P50 NS091856, P30 DK089503, P30AG053760
NSF: 1734853, 1636840, 1416953, 0716055, 1023115
The Elsie Andresen Fiske Research Fund

http://SOCR.umich.edu

Collaborators
• SOCR: Alexandr Kalinin, Selvam Palanimalai, Syed Husain, Matt Leventhal, Ashwini Khare, Rami Elkest, Abhishek Chowdhury, Patrick Tan, Gary Chan, Andy Foglia, Pratyush Pati, Brian Zhang, Juana Sanchez, Dennis Pearl, Kyle Siegist, Rob Gould, Jingshu Xu, Nellie Ponarul, Ming Tang, Asiyah Lin, Nicolas Christou, Hanbo Sun, Tuo Wang
• LONI/INI: Arthur Toga, Roger Woods, Jack Van Horn, Zhuowen Tu, Yonggang Shi, David Shattuck, Elizabeth Sowell, Katherine Narr, Anand Joshi, Shantanu Joshi, Paul Thompson, Luminita Vese, Stan Osher, Stefano Soatto, Seok Moon, Junning Li, Young Sung, Carl Kesselman, Fabio Macciardi, Federica Torri
• UMich MIDAS/MNORC/AD/PD Centers: Cathie Spino, Chuck Burant, Ben Hampstead, Stephen Goutman, Stephen Strobbe, Hiroko Dodge, Hank Paulson, Bill Dauer, Brian Athey

Demo(s)?
• Complex DB Search, retrieval (IDA)
• Multidimensional data visualization (MotionCharts, BrainViewer, R)
• Distributed high-throughput pipeline workflow computing
• SOCRAT Framework
• Data Dashboard
• Education and Training Resources
  • Probability and Statistics Ebook (EBook)
  • Scientific Methods for Health Sciences (SMHS)
  • Data Science and Predictive Analytics (DSPA) MOOC
  • SOCR Tools (distribution calculators, charts, modeler, analyses, experiments)
• Compressive Big Data Analytics (CBDA)