Open Data Science & 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

Outline

- Driving biomedical & health challenges
- Common characteristics of Big Neuroscience Data
- Data science & predictive neuro-analytics methods
- Case-studies
  - Prenatal Exposure to Methamphetamine & Alcohol
  - Parkinson’s disease (PD)
  - UK Biobank (Hands-on Demo)
Population/Census
Unobservable

Big Data
Harmonize/Aggregate Problems

Sample Data
Limited process view

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.
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>Time</td>
<td>Techniques accounting for longitudinal patterns in the data</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

Data Science & Predictive Analytics

- **Data Science**: an emerging extremely transdisciplinary field - bridging between the theoretical, computational, experimental, and applied 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 distributed 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)

http://DSPA.predictive.space

Data Science & Predictive Analytics

- Dimensionality Reduction
- Lazy Learning: Classification Using Nearest Neighbors
- Probabilistic Learning: Classification Using Naive Bayes
- Decision Tree Divide and Conquer Classification
- Forecasting Numeric Data Using Regression Models
- Black Box Machine-Learning Methods: Neural Nets/Support Vector Machines
- Apriori Association Rules Learning
- k-Means Clustering
- Model Performance Assessment
- Improving Model Performance
- Specialized Machine Learning Topics
- Variable/Feature Selection
- Regularized Linear Modeling and Controlled Variable Selection
- Big Longitudinal Data Analysis
- Natural Language Processing/Text Mining
- Prediction and Internal Statistical Cross Validation
- Deep Learning, Neural Networks


Big Data  Information  Knowledge  Action

<table>
<thead>
<tr>
<th>Raw Observations</th>
<th>Processed Data</th>
<th>Maps, Models</th>
<th>Actionable Decisions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Aggregation</td>
<td>Data Fusion</td>
<td>Causal Inference</td>
<td>Treatment Regimens</td>
</tr>
<tr>
<td>Data Scrubbing</td>
<td>Summary Stats</td>
<td>Networks, Analytics</td>
<td>Forecasts, Predictions</td>
</tr>
<tr>
<td>Semantic-Mapping</td>
<td>Derived Biomarkers</td>
<td>Linkages, Associations</td>
<td>Healthcare Outcomes</td>
</tr>
</tbody>
</table>

Dinov, GigaScience (2016) PMID:26918190
Case-Studies – Prenatal Exposure to Methamphetamine & Alcohol

- **Goals:** Examine the local brain effects of prenatal exposure to methamphetamine (MA).
- **Data:** structural magnetic resonance imaging (sMRI). Compared local brain volumes differed among 61 children (ages 5–15 years),
  - 21 with prenatal MA exposure,
  - 18 with concomitant prenatal alcohol exposure (the MAA group),
  - 13 with heavy prenatal alcohol but not MA exposure (ALC group), and
  - 27 unexposed controls.

- **Methods:** Brain morphometry (sMRI processing) & Discriminant analysis (prediction)

- **Results:**
  - Bilateral volume reductions in both exposure groups relative to controls in striatal and thalamic regions, right prefrontal and left occipitoparietal cortices.
  - MAA group had negative correlation between full-scale intelligence quotient (FSIQ) scores and caudate volume.
  - LDA prediction of group membership correctly classified 72% of participants.

- **Conclusions:** Striatal and limbic structures, known to be sites of neurotoxicity in adult MA abusers, may be more vulnerable to prenatal MA exposure than alcohol exposure; Severe striatal damage is associated with more severe cognitive deficit.

*Sowell, et al. JNeurosci (2010)*

---

**Case-Studies – Prenatal Exposure to Methamphetamine & Alcohol**

*p* maps representing the following contrast categories between groups: (1) CON < MAA (red), (2) CON < MAA and CON < ALC (orange), (3) CON < MAA and ALC < MAA (yellow), (4) CON < ALC, CON < MAA, and ALC < MAA (green), (5) CON > MAA and ALC > MAA (dark blue), (6) CON > MAA (light blue), (7) CON > ALC and CON > MAA (purple), and (8) CON > ALC, CON > MAA, and ALC < MAA (light pink).

Factor analysis using Jacobian values for 14 ROIs.

*Sowell, et al. JNeurosci (2010)*
Case-Studies – Parkinson’s Disease

- **Investigate falls in PD patients** using clinical, demographic and neuroimaging data from two independent initiatives (UMich & Tel Aviv U)
- Applied **controlled feature selection** to identify the most salient predictors of patient falls (gait speed, Hoehn and Yahr stage, postural instability and gait difficulty-related measurements)
- **Model-based** (e.g., GLM) and **model-free** (RF, SVM, Xgboost) analytical methods used to forecasts clinical outcomes (e.g., falls)
- Internal statistical cross **validation** + external out-of-bag validation
- Four specific **challenges**
  - Challenge 1, harmonize & aggregate complex, multisource, multisite PD data
  - Challenge 2, identify salient predictive features associated with specific clinical traits, e.g., patient falls
  - Challenge 3, forecast patient falls and evaluate the classification performance
  - Challenge 4, predict tremor dominance (TD) vs. posture instability and gait difficulty (PIGD).
- **Results**: model-free machine learning based techniques provide a more reliable clinical outcome forecasting, e.g., falls in Parkinson’s patients, with classification accuracy of about 70-80%.


---

Case-Studies – Parkinson’s Disease

Falls in PD are extremely difficult to predict …

PD phenotypes
Tremor-Dominant (TD)
Postural Instability & gait difficulty (PI & GD)
## Case-Studies – Parkinson’s Disease

<table>
<thead>
<tr>
<th>Method</th>
<th>acc</th>
<th>sens</th>
<th>spec</th>
<th>ppv</th>
<th>npv</th>
<th>lor</th>
<th>auc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</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>0.774</td>
</tr>
<tr>
<td>Random Forests</td>
<td><strong>0.796</strong></td>
<td><strong>0.683</strong></td>
<td><strong>0.871</strong></td>
<td><strong>0.778</strong></td>
<td><strong>0.806</strong></td>
<td>2.677</td>
<td><strong>0.821</strong></td>
</tr>
<tr>
<td>AdaBoost</td>
<td>0.689</td>
<td>0.610</td>
<td>0.742</td>
<td>0.610</td>
<td>0.742</td>
<td>1.502</td>
<td>0.793</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>0.787</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>0.822</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></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></td>
</tr>
</tbody>
</table>

Results of binary fall/no-fall classification (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)

---

## 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
## Case-Studies – General Populations

<table>
<thead>
<tr>
<th>Case</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>UK Biobank</td>
<td>Discriminate between HC, single and multiple comorbid conditions</td>
</tr>
<tr>
<td></td>
<td>Predict likelihoods of various developmental or aging disorders</td>
</tr>
<tr>
<td></td>
<td>Forecast cancer</td>
</tr>
</tbody>
</table>

### Data Source

<table>
<thead>
<tr>
<th>Sample Size/Data Type</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>UK Biobank</td>
<td>Demographics: &gt; 500K cases</td>
</tr>
<tr>
<td></td>
<td>Clinical data: 4K features</td>
</tr>
<tr>
<td></td>
<td>Imaging data: T1, resting-state fMRI, task fMRI, T2 FLAIR, dMRI, SWI</td>
</tr>
<tr>
<td></td>
<td>Genetics data</td>
</tr>
<tr>
<td></td>
<td>The longitudinal archive of the UK population (NHS)</td>
</tr>
</tbody>
</table>

### Case Studies – UK Biobank

With your own data!

- [http://www.ukbiobank.ac.uk](http://www.ukbiobank.ac.uk)
- [http://bd2k.org](http://bd2k.org)
- [http://myumi.ch/6wQgv](http://myumi.ch/6wQgv)

---

Try-It-Now

With your own data!

- [http://myumi.ch/6wQgv](http://myumi.ch/6wQgv)
- [http://socr.umich.edu/HTML5/SOCR_TensorBoard_UKBB/](http://socr.umich.edu/HTML5/SOCR_TensorBoard_UKBB/)
Acknowledgments

Funding
NIH: P20 NR015331, U54 EB020406, P50 NS091856, P30 DK089503, P30AG053760, UL1TR002240
NSF: 1734853, 1636840, 1416953, 0716055, 1023115
The Elsie Andresen Fiske Research Fund

Collaborators
- SOCR: Alexandr Kalinin, Selvam Palanimalai, Syed Hussain, Matt Leventhal, Ashwini Khare, Ramon Elkeert, Abhishek Chowdhury, Patrick Yan, Gary Chen, Andy Faglia, Pratyush Pali, Brian Zhang, Juana Sanchez, Dennis Pearl, Kyle Siegrist, Rob Gould, Jereme Choi, Yuming Sun, Yi Zhao, Nina Zhou, Nicolas Christou, Hanbo Sun, Tus Wang, Simeone Marino
- LONI/INI: Arthur Toga, Roger Woods, Jack Van Horn, Zhusewei Tu, Yonggang Shi, David Shattuck, Elizabeth Sowell, Katherine Narr, Arand Joshi, Shantenu Joshi, Paul Thompson, Luminita Vese, Stan Osher, Stefano Soatto, Sock Moen, Junrong Li, Young Sung, Carl Kesselman, Fabio Macciardi, Federica Torri
- UMich MIDAS/MNORC/AD/PD Centers: Cathie Spino, Chuck Burant, Ben Hampstead, Stephen Gautman, Stephen Strobbe, Hendrik Dodge, Hank Paulson, Bill Doster, Brian Altroy

http://SOCR.umich.edu