Data Mining and Modeling Methods for Site Inspection Selection

Elena Rantou, Ph.D.
Paul Schuette, Ph.D.

September 14, 2018
Disclaimer

This presentation reflects the views of the presenter and should not be construed to represent the United States Food and Drug Administration’s views or policies.
Outline

- Motivation
- Objectives and background
- Data sets and structures
- Challenges
- Methods and their performance
- Other considerations
Motivation

In a clinical trial setting, data reliability can be jeopardized by:
- Poorly Collected data
- Poorly Processed data
- Poorly Reported data
- Tampered or Fraudulent data
The number and complexity of clinical trials have risen dramatically making it difficult for regulators to choose clinical sites for inspection
Objectives

To determine whether

- supervised data mining methods can be used to predict site inspection results
- unsupervised statistical monitoring can be used to identify ‘unusual’ clinical sites for inspection (*ongoing work*)
Objectives

Onsite inspections help ensure the integrity of the clinical trials via source data verification.

Due to limited resources only less than 1% of the sites can be inspected annually. It is therefore crucial to select the appropriate clinical sites.
Data sets

Site inspection results can be classified into:

- NAI (No Action Indicated)
- VAI (Voluntary Action Indicated)
- OAI (Official Action Indicated)
Data sets

Clinical trial data and the results from clinical site Inspections
Response can be:
- Ordinal with three distinct classes (OAI, VAI, NAI)
- Binary: 2 of 3 ordinal classes are suppressed to 1 (VAI, OAI) vs. NAI
Challenges (ordinal response)

Missing data

Assumptions: missing values are MAR and can be predicted by observed values

Random Forest (RF) imputation

- Replace missing values with sample median
- Use RF to compute proximity between missing and non-missing samples
- Repeat

<table>
<thead>
<tr>
<th>Variable</th>
<th>Type</th>
<th>% missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enrollment</td>
<td>continuous</td>
<td></td>
</tr>
<tr>
<td>Site Specific Efficacy</td>
<td>continuous</td>
<td>27.7%</td>
</tr>
<tr>
<td>Protocol deviation</td>
<td>continuous</td>
<td></td>
</tr>
<tr>
<td>NS adverse event</td>
<td>continuous</td>
<td></td>
</tr>
<tr>
<td>% subject death</td>
<td>continuous</td>
<td></td>
</tr>
<tr>
<td>Enroll/Screen %</td>
<td>continuous</td>
<td></td>
</tr>
<tr>
<td>Subject discontinuation</td>
<td>continuous</td>
<td></td>
</tr>
<tr>
<td>Number of INDs</td>
<td>continuous</td>
<td></td>
</tr>
<tr>
<td>Financial disclosure</td>
<td>continuous</td>
<td>29.9%</td>
</tr>
<tr>
<td>Complaint history</td>
<td>Binary</td>
<td></td>
</tr>
<tr>
<td>Time since last inspection</td>
<td>continuous</td>
<td>4.32%</td>
</tr>
<tr>
<td>OAI history</td>
<td>Binary</td>
<td></td>
</tr>
</tbody>
</table>
Challenges (ordinal response)

Imbalanced outcomes-OAI classification is a rare event with only 1% of sites being classified as OAI.
Challenges (ordinal response)

*Synthetic Minority Over-Sampling Technique-SMOTE*

- Generate synthetic samples for the minority class
- Input the number of nearest neighbors, $k$, $T$ minority class samples and size of SMOTE, $N$
- Output is the synthetic minority class samples
Statistical methods (ordinal response)

- Ordinal regression
- Combined binary classifiers
- Random forests
- Boosted trees
Combined binary classifier

Convert an ordinal regression problem into nested binary classification problems by splitting the data into groups $Y_i \leq j$ and $Y_i > j$ and a binary probability classifier to estimate the probabilities $P(Y_i \leq j)$ and $P(Y_i > j)$.
Classifier performance

OS-OE curve generated by threshold

OS-OE curve generated by SMOTE

OS-OE curve generated by SMOTE
Statistical methods (binary response)

- Random Forest
- Boosted Tree
- Boosted Dropout

(As boosting is susceptible to overfitting-high bias, low variance)
Challenges (binary response)

- Studying the sensitivity of each variable to predict the outcome
- Using the EM-algorithm to impute missing data
- Using 5-fold cross-validation to assess model performance
Classifier Performance

ROC curve of Methods

- True positive rate vs. False positive rate
- Colors represent different methods:
  - Random Forest
  - Boosted Tree
  - Boosted Dropout
## Model performance

<table>
<thead>
<tr>
<th>Method</th>
<th>CV error</th>
<th>Misclassification</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>13.5%</td>
<td>14.0%</td>
</tr>
<tr>
<td>Boosted Tree</td>
<td>15.9%</td>
<td>14.9%</td>
</tr>
<tr>
<td>Boosted Tree with Dropout</td>
<td>16.9%</td>
<td>16.4%</td>
</tr>
</tbody>
</table>
Outcome

R-Shiny application that uses the supervised learning methods and

- Predicts the potentially fraudulent cases from different clinical sites
- Validates the parameter that gives the best fit
- Detects the covariates that are most predictive of the outcomes
CRADA

Cooperative Research and Development Agreement with CluePoints

The main objective is to detect atypical sites in a multicenter study

Method tests the distribution of data in one center with data in other centers and produces a p-value demonstrating how unlikely the outcomes from one clinical center are (unsupervised approach)

Approaching the end of 2\textsuperscript{nd} year is a 3 year agreement
References


Acknowledgements

Mingwei Tang
Chetkar Jha
Nicholas Hein
Thank you!