Patient Risk Prediction Model via Top-$k$ Stability Selection

Jiayu Zhou$^1$, Jimeng Sun$^2$, Yashu Liu$^1$, Jianying Hu$^2$, Jieping Ye$^1$

$^1$Arizona State University, AZ
$^2$IBM T.J. Watson Research Center, NY
Outline

• Sparse learning and model selection
• Stability selection
• Top-k stability selection
• Patient risk prediction via Top-k Stability Selection
Outline

• Sparse learning and model selection
• Stability selection
• Top-k stability selection
• Patient risk prediction via Top-k Stability Selection
Sparse Learning

• Many data mining problems can be viewed as obtaining model parameters, represented by a vector or a matrix.

• Take linear models as an example: each model parameter corresponds to a feature

\[
f(\text{Sample size}) \times \text{Dimension} = \text{Data Matrix} \times \text{Dimension} \Rightarrow \text{Prediction}
\]

\[
w = \arg\min_w L(f(Xw), y)
\]
Sparse Learning

• Sparsity is essentially a way to induce “simplicity”, that many of the model parameters are zeros
  – A feature is considered to be irrelevant if the corresponding parameter is zero.

• Benefits
  – Model interpretability
  – Generalization performance
  – Simultaneous feature selection and predictive modeling
Sparse Learning

• Typically sparsity in model is achieved by introducing sparse-inducing norms in the objective

\[
\min_w L(w) + \lambda R(w)
\]

• Model selection
  – Determine a proper amount of regularization
    • A simple model with a few features
      • Include the features related to the prediction problem.
  – Typically done by cross validation
  – Very sensitive to data perturbation
Outline

• Sparse learning and model selection

• **Stability selection**

• Top-k stability selection

• Patient risk prediction via Top-k Stability Selection
Stability Selection

• Address the problem of model selection using resampling techniques.

• A formal framework of model selection with certain empirical and theoretical advantages.
  – Guarantees under certain assumptions a certain bound on the expected number of false selections given a finite sample size

• The choice of the initial set of regularization parameters typically does not have a very strong influence on the results
Stability Selection: The Procedure

• Step 1: Computing selection probability

Sparse Learning Solver with Parameter $\lambda = 0.04$

Bootstrap Datasets

Models Returned by Solver

Selection Probability $\lambda = 0.04$

$Pr = 1$
$Pr = 0.24$
$Pr = 0.78$
$Pr = 0.55$
$Pr = 0$
$Pr = 0.44$
$Pr = 0.52$
Stability Path VS Lasso Path

- Gene expression data with 4088 genes.
- Choosing the right parameter is less critical for stability path

Figure from Meinshausen and Buhlmann 2010
Stability Selection: The Procedure

- Step 2: Computing stable set

\[ \Lambda(1) = 0.04 \]  \[ \Lambda(2) = 0.08 \]  \[ \Lambda(3) = 0.16 \]  \[ \Lambda(4) = 0.80 \]
Computational Cost of Stability Selection

• The computation on each bootstrap dataset is independent
  – Efficient path-wise warm-start strategy
  – Recently developed (sequential) screening technique
  – Parallel computation for each bootstrap dataset

• The choice of range of parameters and number of lambda
  – Use bisection on the entire data to find a parameter range with proper number of features.
  – Stable performance with 15 or larger values of lambda.
Applications of Stability Selection

• Identify risk factors for childhood malnutrition
  – Febske et al., Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression, JASA, 2011

• Discovery molecular signatures
  – Haury et al., The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures, PLoS ONE, 2011

• Genome wide association studies (GWAS)
  – He et al., A variable selection method for genome-wide association studies, Bioinformatics, 2011

• Identify biomarkers for Alzheimer’s disease
Outline

• Sparse learning and model selection
• Stability selection
• Top-k stability selection
• Patient risk prediction via Top-k Stability Selection
Stability Selection

• MAX operator: what if some features only perform well at one parameter?

\[ \Lambda_{(1)} = 0.04 \]
\[ \Lambda_{(2)} = 0.08 \]
\[ \Lambda_{(3)} = 0.16 \]
\[ \Lambda_{(t)} = 0.80 \]
Our Approach: Top-\(k\) Stability Selection

- Why not use more information?

\[
\begin{align*}
\Lambda_{(1)} &= 0.04 \\
\Lambda_{(2)} &= 0.08 \\
\Lambda_{(3)} &= 0.16 \\
\Lambda_{(t)} &= 0.80
\end{align*}
\]

Average of Top \(k\) Selection Probability

\[
\begin{align*}
\Lambda &= 0.99 \\
\Lambda &= 0.24 \\
\Lambda &= 0.52 \\
\Lambda &= 0.79 \\
\Lambda &= 0.1 \\
\Lambda &= 0.52 \\
\Lambda &= 0.44
\end{align*}
\]

Pr = 0.99
Pr = 0.24
Pr = 0.52
Pr = 0.79
Pr = 0.1
Pr = 0.52
Pr = 0.44

Threshold \( \Pr \geq 0.5 \)

Stable Score \( \Lambda \)

Stable Set \( \Lambda \)
Performance of Top-k Stability Selection

- Experimental results on the synthetic data:
  - Top-k stability selection delivers significantly better performance than original stability selection
  - A higher $k$ value typically gives a better result.
Why Top-\(k\) Stability is Better

• Error control property of top-\(k\) stability selection:

\[ \frac{\mathbb{E}\left(\left|S \cap \hat{S}^\lambda\right|\right)}{|S|} \geq \frac{\mathbb{E}\left(\left|N \cap \hat{S}^\lambda\right|\right)}{|N|}, \]

and for \(\forall \lambda \in \Lambda\) the distribution of \(\left\{\mathbf{1}\{f \in \hat{S}^\lambda\}, f \in N\right\}\) is exchangeable, then,

\[ \mathbb{E}(V_k) \leq \frac{\sum_{i=1}^{k} u_{\Lambda,i}^2}{k \cdot p \cdot (2\pi_{th} - 1)}, \]

where \(\frac{1}{2} < \pi_{th} < 1\), \(u_{\Lambda,i}^2 = \mathbb{E}[\Lambda_f^{-i}]^2\), \(i = 1, 2, \ldots, k\), \(\Lambda_f^{-i}\) is defined the same as in the definition of top-\(k\) stable features.

• Key insights:
  
  – The upper bound of error in top-\(k\) stability selection is at least as tight as the original stability selection
  
  – By utilizing the information from top-\(k\) stability score, the top-\(k\) approach is more robust.
Outline

• Sparse learning and model selection
• Stability selection
• Top-k stability selection
• Patient risk prediction via Top-k Stability Selection
Patient Risk Prediction from Electronic Health Record

• Electronic health records (EHRs) capture longitudinal patient information in digital forms from diverse sources
  – Construct features of risk prediction
  – Identify high quality features and biomarkers
  – Rank all available features using feature selection methods
Heart Failure Risk Prediction

• Heart failure (HF) is currently the leading cause of hospitalization among Medicare beneficiaries.

• Framingham risk criteria, being the most commonly used diagnostic criteria for HF
  – Not systematically documented in structured EHR
  – Requires natural language processing to extract, which is expensive and time-consuming.

• Goal: construct features from structured EHRs, and use a few effective features to predict Framingham criteria.
Feature Construction from EHR

- EHR data documents detailed patient encounters (e.g., diagnosis, medication, lab result, clinical notes)
- Goal: capture sufficient clinical nuances for risk prediction task.
- Summarize (aggregate) events into feature variables
Experiment Results

- Predictive performance on EHR data
  - Higher k value gives a better predictive
  - Not the case that the larger k the better

Table 2: Average AUC ranks of feature selection algorithms over the 9 datasets.

<table>
<thead>
<tr>
<th>Feature #</th>
<th>4</th>
<th>6</th>
<th>8</th>
<th>10</th>
<th>12</th>
<th>14</th>
<th>16</th>
<th>18</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>ReliefF</td>
<td>10.0 ± 0.0</td>
<td>10.0 ± 0.0</td>
<td>10.0 ± 0.0</td>
<td>10.0 ± 0.0</td>
<td>10.0 ± 0.0</td>
<td>10.0 ± 0.0</td>
<td>10.0 ± 0.0</td>
<td>10.0 ± 0.0</td>
<td>10.0 ± 0.0</td>
</tr>
<tr>
<td>Fisher</td>
<td>7.0 ± 2.1</td>
<td>7.2 ± 1.5</td>
<td>6.2 ± 2.4</td>
<td>6.0 ± 2.6</td>
<td>5.8 ± 2.6</td>
<td>5.6 ± 2.2</td>
<td>4.9 ± 2.8</td>
<td>5.0 ± 2.1</td>
<td>5.6 ± 1.5</td>
</tr>
<tr>
<td>Gini</td>
<td>7.0 ± 1.5</td>
<td>7.1 ± 2.0</td>
<td>6.0 ± 1.8</td>
<td>6.8 ± 1.8</td>
<td>6.9 ± 2.1</td>
<td>6.4 ± 2.8</td>
<td>5.9 ± 2.0</td>
<td>6.6 ± 2.0</td>
<td>6.1 ± 2.1</td>
</tr>
<tr>
<td>InfoGain</td>
<td>4.7 ± 2.3</td>
<td>6.2 ± 2.0</td>
<td>5.7 ± 1.7</td>
<td>5.7 ± 1.7</td>
<td>5.8 ± 1.1</td>
<td>5.9 ± 1.3</td>
<td>6.0 ± 2.4</td>
<td>7.0 ± 1.6</td>
<td>6.8 ± 1.7</td>
</tr>
<tr>
<td>ChiSquare</td>
<td>5.3 ± 2.2</td>
<td>6.8 ± 1.8</td>
<td>6.4 ± 2.5</td>
<td>6.6 ± 1.6</td>
<td>6.2 ± 1.4</td>
<td>5.8 ± 2.0</td>
<td>6.7 ± 1.4</td>
<td>7.2 ± 1.5</td>
<td>7.2 ± 1.6</td>
</tr>
<tr>
<td>mRMR</td>
<td>5.8 ± 2.6</td>
<td>5.7 ± 1.7</td>
<td>7.1 ± 1.5</td>
<td>7.3 ± 1.7</td>
<td>7.8 ± 1.5</td>
<td>7.6 ± 1.6</td>
<td>7.2 ± 1.7</td>
<td>6.9 ± 2.4</td>
<td>6.4 ± 3.4</td>
</tr>
<tr>
<td>SS(k-1)</td>
<td>5.1 ± 3.0</td>
<td>4.1 ± 2.2</td>
<td>4.8 ± 2.5</td>
<td>4.4 ± 2.5</td>
<td>4.6 ± 2.7</td>
<td>4.2 ± 3.6</td>
<td>4.7 ± 3.2</td>
<td>3.8 ± 2.4</td>
<td>4.8 ± 2.3</td>
</tr>
<tr>
<td>SS(k-2)</td>
<td>3.8 ± 3.0</td>
<td>3.0 ± 2.4</td>
<td>2.8 ± 2.0</td>
<td>2.4 ± 1.9</td>
<td>3.2 ± 2.0</td>
<td>3.7 ± 2.1</td>
<td>3.2 ± 2.7</td>
<td>3.3 ± 1.9</td>
<td>3.0 ± 1.7</td>
</tr>
<tr>
<td>SS(k-4)</td>
<td>3.0 ± 1.8</td>
<td>2.3 ± 1.6</td>
<td>2.8 ± 2.3</td>
<td>2.2 ± 1.6</td>
<td>2.6 ± 1.6</td>
<td>2.8 ± 1.6</td>
<td>2.9 ± 1.4</td>
<td>2.3 ± 1.3</td>
<td>2.3 ± 1.6</td>
</tr>
<tr>
<td>SS(k-8)</td>
<td>3.3 ± 1.9</td>
<td>2.6 ± 1.2</td>
<td>3.2 ± 2.4</td>
<td>3.6 ± 1.9</td>
<td>2.2 ± 1.5</td>
<td>3.1 ± 1.8</td>
<td>3.6 ± 1.9</td>
<td>2.9 ± 1.7</td>
<td>2.8 ± 1.8</td>
</tr>
</tbody>
</table>
Experimental Results

• Comparison with sparse logistic regression
  – When selecting the same number of features, the top-k stability selection outperforms sparse logistic regression in terms of predictive performance.
Experiment Results

• Experiments on EHR data
  – The rank of important features improves as $k$ increases.
  – ‘Cut-off’ points increase – easy to choose a threshold.
  – Trade-off between stability and sensitivity
Conclusion and Future Works

• Conclusion
  – Proposed top-k stability selection and established its theoretical bound on error control
  – Demonstrated the effectiveness of top-k stability selection on both synthetic and EHR data

• Future works
  – Top-k stability selection is based on Lasso and thus is vulnerable to strongly correlated features
  – Apply top-k stability selection and structured sparsity to explore the EHR
Thanks!