An Integrated Machine Learning Approach to Stroke Prediction

Aditya Khosla     Yu Cao     Cliff Chiung-Yu Lin
Hsu-Kuang Chiu    Junling Hu*     Honglak Lee

Stanford University

*eBay Inc. (formerly at Robert Bosch Corporation)
Outline

- Motivation
- Our Approach
  - Data imputation, feature selection, and prediction
  - A new algorithm for feature selection
  - A new algorithm for prediction
- Experimental Results
- Summary
Motivation
Importance of stroke prediction

- The third leading cause of death in the US
  - 137,000 die from stroke each year.
- Leading cause of long-term disability in the US
- Risk factors need to be discovered.
- Current research on stroke is on simple statistical models.

Our goal: Bring machine learning methods to stroke prediction.
Identifying risk factors

- Mostly based on clinical studies
- Known risk factors
  - Physical:
    - E.g.: Age, prior stroke, blood pressure, hypertension, time to walk 15 feet, cardiac injury score, diabetic status, atrial fibrillation, left ventricular mass, etc.
  - Behavioral:
    - E.g.: cigarette smoking, poor diet, alcohol abuse, etc.
Existing stroke prediction models

- Cox proportional hazards model
  - One of the most commonly used statistical methods in medical research
  - Applied to prediction of various diseases

Hazard function at time $t$

$$h(t \mid \mathbf{x}; \beta) = h_0(t) \exp(\beta^T \mathbf{x})$$

- $\mathbf{x}$: input features for an individual
- $t$: timing of stroke
- $\beta$: parameters of the model
Previous approaches

- Related work on stroke prediction
  - Lumley et al. (2002), Manolio et al. (1996); Longstreth et al. (2001); Chambless et al. (2004); Hitman et al. (2007), etc.

- Limitations
  - Use limited number of features
    - Manually selected
    - Small size (< 20)
  - Limited modeling methods
    - Most used Cox proportional hazards regression
    - Not utilizing modern machine learning methods
Our Approach
## Existing approaches vs. Our approach

<table>
<thead>
<tr>
<th></th>
<th>Existing approaches</th>
<th>Our approach</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of features</strong></td>
<td>~ 20</td>
<td>~ 1000</td>
</tr>
<tr>
<td><strong>Feature selection</strong></td>
<td>Manually selected</td>
<td>Automatic feature selection</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(e.g., L1 logistic regression)</td>
</tr>
<tr>
<td><strong>Prediction algorithm</strong></td>
<td>Cox proportional hazards model</td>
<td>Machine learning methods</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(e.g., SVM)</td>
</tr>
</tbody>
</table>

**Examples of existing approaches:**
Lumley et al. (2002); Manolio et al. (1996); Longstreth et al. (2001); Chambless et al. (2004); Hitman et al. (2007), etc.
Overview of our approach

Data Imputation
- “Mean”
- “Median”
- Linear regression
- ...

Feature selection
- L1 logistic regression
- Conservative mean feature selection
- ...

Prediction
- SVM
- Margin-based Censored regression
- ...

Meanings of Symbols:
- X: Methods are not applicable or not used.
- Arrow: Method is used.
Our methods

- We evaluated several missing value imputation methods
  - Mean, median, linear regression, EM.
- We evaluated several feature selection methods
  - Forward feature selection
  - L1-regularized logistic regression
  - Conservative Mean feature selection (this paper)
- We evaluated several prediction methods
  - SVM (*SVM-perf* to directly optimize the AUC)
  - Margin-based Censored regression (this paper)
Feature selection: Conservative Mean

- For each feature \( j \), divide the training data into \( N \) folds and compute:
  
  \[
  AUC^k : \text{Area under the ROC curve for fold } k
  \]
  
  \[
  \mu_j = \frac{1}{N} \sum_{k=1}^{N} AUC^k
  \]
  
  \[
  \sigma_j = \sqrt{\frac{1}{N} \sum_{k=1}^{N} (AUC^k - \mu_j)^2}
  \]

- Use \( \mu_j - \sigma_j \) for ranking the features (i.e., more “conservative” estimate than \( \mu_j \)).
  - Details in the paper.
Margin-based Censored Regression (MCR)

- **Prediction function**
  - Want to learn: \( z \sim w^T x \)

- **Censored regression**
  - Want to predict timing of stroke only if it happens within a given timeframe.

- **“Margin-based”**
  - If stroke does not happen, we want to predict it as “negative” with a margin.

\[
\begin{align*}
    &x: \text{features} \\
    &z: \text{“inverse” of stroke timing} \\
    &z > 0: \text{stroke happened} \\
    &z \leq 0: \text{stroke did not happen}
\end{align*}
\]
Optimization problem for MCR

- We solve the following optimization problem:

\[
\begin{align*}
\text{minimize} & \quad w, \xi \\
& \quad \sum_{i:y(i)=1} \phi(w^T x^{(i)} - z(\tilde{t}^{(i)})) + C \sum_{i:y(i)=0} \xi^{(i)} + \gamma ||w||_2^2 \\
\text{subject to} & \quad w^T x^{(i)} \leq -\epsilon + \xi^{(i)}, \quad \forall i \in \{i | y^{(i)} = 0\}, \\
& \quad \xi^{(i)} \geq 0, \quad \forall i,
\end{align*}
\]

regression error for stroke events
classification error for "non-stroke" cases
margin constraints
Experimental results
Experimental setup

- Cardiovascular Heart Study (CHS) data
  - Annual examinations for elderly people (+65 years)
  - Study conducted from 1989 for 10+ years
- After preprocessing, we have 796 features, 4988 examples (299 positives/ 4689 negatives)
- Our task
  - Use baseline (first year) measurement as features and perform 5 year prediction
  - Train over 9/10 of data and test on 1/10 of data (random split and repeat 5 times).
Results – missing data imputation

- Used Conservative Mean for feature selection and SVM for prediction.
  - For each missing value, substituting with the median (over the observed feature values) performed the best

<table>
<thead>
<tr>
<th>Imputation Method</th>
<th>Test AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Column Median</td>
<td><strong>0.774</strong></td>
</tr>
<tr>
<td>Linear Regression (with rounding)</td>
<td>0.768</td>
</tr>
<tr>
<td>Regularized EM</td>
<td>0.765</td>
</tr>
<tr>
<td>Column Mean (with rounding)</td>
<td>0.765</td>
</tr>
</tbody>
</table>
Prediction results - AUC

- Best performance achieved using Conservative mean + MCR
  - 15% error reduction over Lumley et al.’s method

<table>
<thead>
<tr>
<th>Test AUC</th>
<th>Prediction algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature selection algorithm</td>
<td>SVM</td>
</tr>
<tr>
<td>Conservative Mean</td>
<td>0.774</td>
</tr>
<tr>
<td>L1 logistic regression</td>
<td>0.764</td>
</tr>
<tr>
<td>Manually selected 16 features*</td>
<td>0.753</td>
</tr>
</tbody>
</table>

**Baseline:** Cox + 16 features*: 0.734

* used in Lumley et al. (2002)
Prediction results – Concordance Index

- Similar results as AUC

<table>
<thead>
<tr>
<th>Test Concordance Index</th>
<th>Prediction algorithm</th>
<th>SVM</th>
<th>MCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature selection algorithm</td>
<td></td>
<td>SVM</td>
<td>MCR</td>
</tr>
<tr>
<td>Conservative Mean</td>
<td>0.760</td>
<td>0.770</td>
<td></td>
</tr>
<tr>
<td>Manually selected 16 features*</td>
<td>0.747</td>
<td>0.757</td>
<td></td>
</tr>
</tbody>
</table>

**Baseline:** Cox + 16 features*: 0.730

* used in Lumley et al. (2002)
Discovering potential risk factors

- Top features selected by our algorithm from a set of 796 features (or measurements)

<table>
<thead>
<tr>
<th>Description</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>0.606</td>
</tr>
<tr>
<td><strong>Number of symbols correctly coded</strong>*</td>
<td>0.583</td>
</tr>
<tr>
<td><strong>Maximal inflation level</strong>*</td>
<td>0.582</td>
</tr>
<tr>
<td>Systolic blood pressure</td>
<td>0.574</td>
</tr>
<tr>
<td><strong>Calculated 100 point score</strong>*</td>
<td>0.568</td>
</tr>
<tr>
<td><strong>Total medications</strong>*</td>
<td>0.563</td>
</tr>
<tr>
<td>Isolated systolic hypertension</td>
<td>0.559</td>
</tr>
<tr>
<td><strong>General health</strong>*</td>
<td>0.552</td>
</tr>
<tr>
<td>Calculated hypertension status</td>
<td>0.550</td>
</tr>
<tr>
<td>Time (in sec) to walk 15 feet</td>
<td>0.549</td>
</tr>
</tbody>
</table>

* These represent newly discovered potential risk factors.
Summary

- Integrated approach to stroke prediction
  - Imputation, feature selection, and prediction
- Novel feature selection/prediction algorithms
  - Conservative Mean feature selection
  - Margin-based Censored Regression
- Outperform the existing methods
- Discovery of new potential risk factors
Thank you!