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

$$\overbrace{h(t \mid \mathbf{x}; \beta)}^{\text{Hazard function at time } t} = h_0(t) \exp(\beta^T \mathbf{x})$$

$\mathbf{x}$: input features for an individual

t : timing of stroke

β : 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

	Existing approaches	Our approach
Number of features	~ 20	~ 1000
Feature selection	Manually selected	Automatic feature selection (e.g., L1 logistic regression)
Prediction algorithm	Cox proportional hazards model	Machine learning methods (e.g., SVM)

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
- …

Our methods

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

Feature selection: Conservative Mean

- For <u>each feature</u> *j*, divide the training data into *N* folds and compute:

$$AUC^k : \text{Area under the ROC curve for fold } k$$

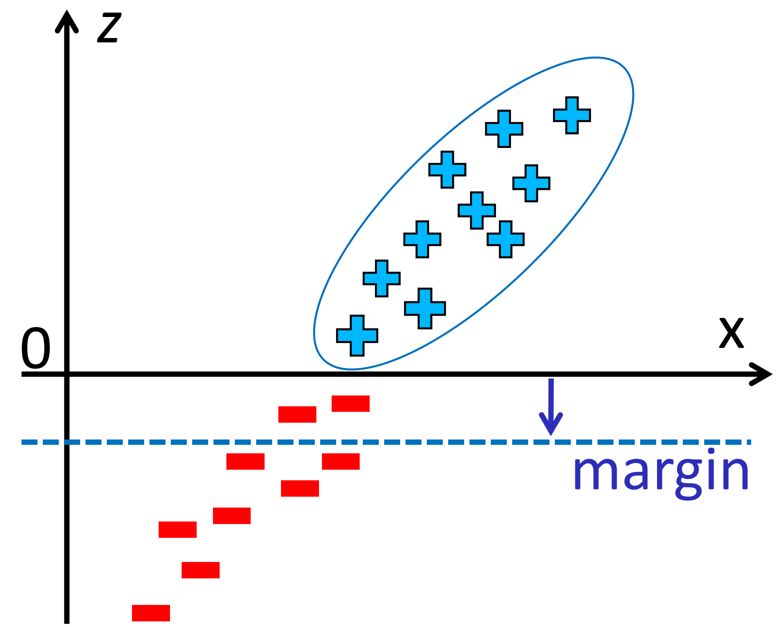
$$\mu_j = \frac{1}{N}\sum_{k=1}^{N} AUC^k$$

$$\sigma_j = \sqrt{\frac{1}{N}\sum_{k=1}^{N} (AUC^k - \mu_j)^2}$$

- Use $\mu_j - \sigma_j$ for ranking the features (i.e., more "conservative" estimate than μ_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.



x: features

z: "inverse" of stroke timing t

- $z > 0$: stroke happened
- $z \leq 0$: stroke did not happen

Optimization problem for MCR

- We solve the following optimization problem:

$$\text{minimize}_{\mathbf{w},\xi} \quad \overbrace{\sum_{i:y^{(i)}=1} \phi(\mathbf{w}^T\mathbf{x}^{(i)} - z(\tilde{t}^{(i)}))}^{\text{regression error for stroke events}} + \overbrace{C \sum_{i:y^{(i)}=0} \xi^{(i)}}^{\text{classification error for "non-stroke" cases}} + \gamma||\mathbf{w}||_2^2$$

$$\text{subject to} \quad \left.\begin{array}{l} \mathbf{w}^T\mathbf{x}^{(i)} \leq -\epsilon + \xi^{(i)}, \ \forall i \in \{i | y^{(i)} = 0\}, \\ \xi^{(i)} \geq 0, \ \forall i, \end{array}\right\} \text{margin constraints}$$

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

Imputation Method	Test AUC
Column Median	**0.774**
Linear Regression (with rounding)	0.768
Regularized EM	0.765
Column Mean (with rounding)	0.765

Prediction results - AUC

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

Test AUC	Prediction algorithm	
Feature selection algorithm	SVM	MCR
Conservative Mean	0.774	**0.777**
L1 logistic regression	0.764	0.771
Manually selected 16 features*	0.753	0.765

Baseline: Cox + 16 features*: 0.734

* used in Lumley et al. (2002)

Prediction results – Concordance Index

- Similar results as AUC

Test Concordance Index	Prediction algorithm	
Feature selection algorithm	SVM	MCR
Conservative Mean	0.760	**0.770**
Manually selected 16 features*	0.747	0.757

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)

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

* 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!