Automated feature selection of predictors in electronic medical records data

Gronsbell, Jessica, Jessica Minnier, Sheng Yu, Katherine Liao, and Tianxi Cai. "Automated feature selection of predictors in electronic medical records data." *Biometrics* 75, no. 1 (2019): 268-277.

Presenter: Yi Han, Yaqi Shi

December 6th, 2022

Agenda

- Introduction & Motivation
- Methods
- Simulation
- Conclusion

Introduction & Motivation

- Electronic Health Record (EHR) contains many clinical data that are useful for identifying disease phenotypes, which can be used in translational research in omics studies
- Two historical approaches to identify disease status:
  - International Classification of Diseases Ninth Edition (ICD-9) billing code

    --imprecise billing code can lead to incorrect prediction of disease phenotypes
  - Gold standard labels (review medical chart)

    -- require significant amount of manual work
- This paper discusses a novel procedure to accurately and efficiently identify disease phenotypes based on EHR data (automatically select features based on unlabeled observations)

Methods - Notation and Assumption

- Data: N iid random vector $\mathscr{F} = \left\{ \left(Y_i, \mathbf{X}_i^\top, \mathbf{S}_i^\top \right)^\top, i = 1, \ldots, N \right\}.$
- Observed set: $\mathscr{D} = \left\{ \mathbf{W}_i = \left(\mathbf{X}_i^\top, \mathbf{S}_i^\top \right)^\top, i = 1, \ldots, N \right\}$
- Assumption:
    - Y follows a GLM $P(Y = 1 \mid \mathbf{X}) = g\left(\alpha_0 + \mathbf{X}^\top \beta_0 \right) = g\left(\vec{\mathbf{X}}^\top \theta_0 \right)$
      $\vec{\mathbf{X}} = \left(1, \mathbf{X}^\top \right)^\top, \ \theta_0 = \left(\alpha_0, \beta_0^\top \right)^\top$
    - X is elliptical symmetric
    - S depends on X through Y $\mathbf{S} \perp \mathbf{X} \mid Y$

Methods

- Step 1: Unsupervised feature selection procedure
- Step 2: Variable selection via resampling

Methods - Unsupervised feature selection procedure

- Two steps: Clustering and Regularized Estimation
- 1. Clsutering: estimate $\pi_{\mathbf{S}} = P(Y = 1 \mid \mathbf{S})$

$$\mathbf{S} \sim \tau f_{\Theta_1}(\mathbf{s}) + (1-\tau) f_{\Theta_0}(\mathbf{s})$$

$$\Pi_{\mathbf{S}}(\Theta_\bullet) = \frac{\tau f_{\Theta_1}(\mathbf{S})}{\tau f_{\Theta_1}(\mathbf{S}) + (1-\tau) f_{\Theta_0}(\mathbf{S})},$$

- 2. Regularized Estimation: find $\widehat{\mathcal{A}} = \{j : \widehat{\beta}_j \neq 0\}$

$$\widehat{\mathcal{L}}(\theta) = N^{-1} \sum_{i=1}^{N} \ell\left(\theta^\top \overrightarrow{\mathbf{X}}_i, \hat{\pi}_{\mathbf{S}_i}\right) + \lambda_N \sum_{j=1}^{p} |\beta_j| / \left|\widetilde{\beta}_j\right|$$

$$\left\|\widetilde{Y} - \theta^\top \widetilde{X}\right\|_2^2 + \lambda_N \sum_{j=1}^{p} |\beta_j| / \left|\widetilde{\beta}_j\right|$$

Methods - Variable selection via resampling

- One time feature selection is not stable
- Subsampling: $\widehat{\mathcal{L}}^{(m)}(\boldsymbol{\theta}) = N_b^{-1} \sum_{i \in \mathcal{R}_m} \ell(\boldsymbol{\theta}^\top \vec{\mathbf{X}}_i, \widehat{\pi}_{\mathbf{S}_i}) + \lambda_N^{(m)} \sum_{j=1}^{p} |\beta_j| / |\widetilde{\beta}_j^{(m)}|$ $\quad \widehat{\boldsymbol{\theta}}^{(m)} = (\widehat{\alpha}^{(m)}, \widehat{\boldsymbol{\beta}}^{(m)\top})^\top$

$$\widehat{\rho}_{0j} = M^{-1} \sum_{m=1}^{M} I\left(\widehat{\beta}_j^{(m)} = 0\right)$$

- Select the feature if the probability is less than the cutoff

  Common choice of the cutoff is 0.5

Simulations under 3 Settings

- Setting 1: all assumptions hold (distribution of X is elliptically symmetric, Y follows a logistic regression model)

  verify algorithm performance under the correct assumptions

- Setting 2: distribution of X is not elliptically symmetric

  determine the robustness of the algorithm when assumption on distribution of X does not hold

- Setting 3: S is not conditionally independent of X given Y

  determine whether the algorithm works when the conditional independence assumption does not hold

Simulation Methods

- In each setting, N=5000 subjects were generated with binary outcome Y with prevalence=0.3
- X and S are generated based on MVN distribution (in setting 2, applied transformation log(exp(X)+1) so that X is not symmetrically distributed; in setting 3, add X to S so that S is not conditionally independent of X)
- Considered p=50 or 100 features
- Compared supervised methods that directly fit Y on X with other existing unsupervised feature selection methods, specifically, we are interested in the methods using the proposed automated clustering procedure with/without resampling (*AutoClust, AutoClust_R*)
- After feature selection, train the final algorithm on 100 or 200 labeled samples, calculate area under the receiver operating characteristic curve (*AUC_100, AUC_200*) to indicate prediction performance
- Repeat 500 times and obtain average estimates

Simulation Results

- In all 3 settings:
- supervised methods tend to produce overly simple models which have weaker prediction performance
- the automated selection procedure has improved the prediction performance compared to directly training a supervised algorithm on the features
- *AutoClust_R* has the highest *AUC_100* and *AUC_200* among all model
- *AUC_100* of *AutoClust_R* is similar to or larger than *AUC_200* of the supervised methods
    - automated clustering feature selection + resampling can achieve similar or better performance than supervised methods using half as many labeled samples
    - Can reduce labor work for gold standard labelling

Conclusion

- New Method - Automated feature selection method with resampling
  - Unsupervised feature selection procedure
  - Variable selection via resampling
- Method Performance
  - AutoClust and AutoClust_R vs Other Algorithms
    - Accurate feature selection
    - High predictive preformance
  - AutoClust vs AutoClust_R
    - Resampling shows some improvement
- Advancement
  - Multiple Non-binary surrogates, high-dimensonal predictor and unlabelled set

Thank you