Classification using Ensemble Learning under Weighted Misclassification Loss

Yizhen Xu

PhD Candidate, Brown University

July 30th, 2017

Motivation - Classification Rule

- Input X, binary output Y
- Based on weighted misclassification loss, develop a classification rule $Q(X)$ that classifies Y
- $Q(X) = Q(X; \Psi(\cdot; \alpha), c) = \mathbb{1}\{\Psi(X; \alpha) \geq c\}$
- Risk score $\Psi(X; \alpha)$, threshold c
- We want to use Super Learner to get a risk score and minimize the weighted misclassification risk (Vaart and Laan, 2006; Laan and Polley, 2007)

Motivation - Examples

- Kenyan clinical HIV data
  - 899 complete cases; derived from three studies conducted at the Academic Model Providing Access to Healthcare (AMPATH) in Eldoret, Kenya (Mann et al. 2013; Diero et al. 2014; Brooks et al. 2016)
  - Y: viral failure (VL > 1000 copies/ml)
  - X: age, gender, nadir CD4, CD4, CD4 percent, adherence to ART, time since starting current ART, and slope of CD4 percent progression
- Wisconsin diagnostic breast cancer data
  - 569 cases; available on UCI data repository
  - Y: confirmatory diagnosis of breast cancer as either benign or malignant
  - X: 30 covariates derived from 10 cell image features

Motivation - Problem

- Most applications weight false positives (FP) and false negatives (FN) equally
- Viral failure classification in HIV treatment monitoring
  - Viral load (VL) assessment may be limited by logistics, cost, and technology
  - Predict viral failure (VL > 1000 copies/ml) based on other clinical markers
  - FP: early treatment switching, higher toxicity, lower adherence, greater costs, limited long term treatment options
  - FN: drug resistence, increased morbidity and mortality
- Weighted misclassification loss: FP and FN are treated differently

Thresholding

- Common approach: conditional thresholding
  - Estimate risk score
  - Set threshold conditional on the estimated risk score
- Our strategy: joint thresholding
  - Simultaneous estimation of risk score and threshold under weighted misclassification loss
  - This joint estimation give more accurate estimate and improvement to overall risk compared to the common approach

Weighted Misclassification Loss (WML)

Recall:

- Rule $Q(X) = \mathbb{1}\{\Psi(X;\alpha) \geq c\}$
- Outcome Y

Loss:

$$L_\lambda(Y, Q(X)) = \lambda \mathbb{1}\{Q(X) = 0, Y = 1\} + (1 - \lambda)\mathbb{1}\{Q(X) = 1, Y = 0\}$$

λ is a user specified weight that governs FP and FN

Risk:

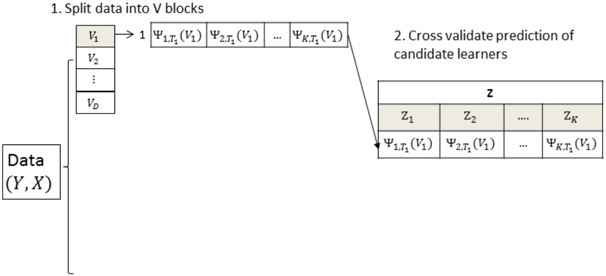
$$R_\lambda(Y, Q(X)) = \lambda P(Q(X) = 0, Y = 1) + (1 - \lambda) P(Q(X) = 1, Y = 0)$$

Empirical Weighted Misclassification Risk

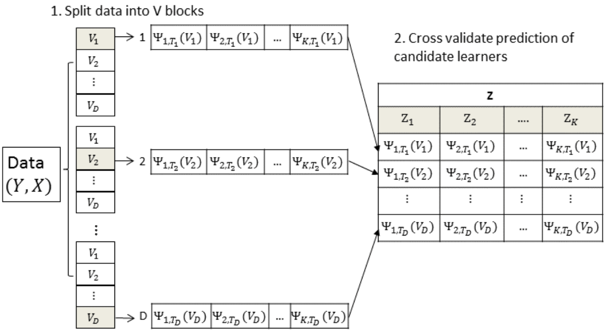
$$\widehat{R}_\lambda(Y, \Psi(X;\alpha), c) = \frac{1}{n}\sum_{i=1}^{n} \lambda \mathbb{1}\{\Psi(X_i;\alpha) < c, Y_i = 1\}$$
$$+ (1-\lambda)\mathbb{1}\{\Psi(X_i;\alpha) \geq c, Y_i = 0\}$$

Find α and c that minimize the empirical risk function

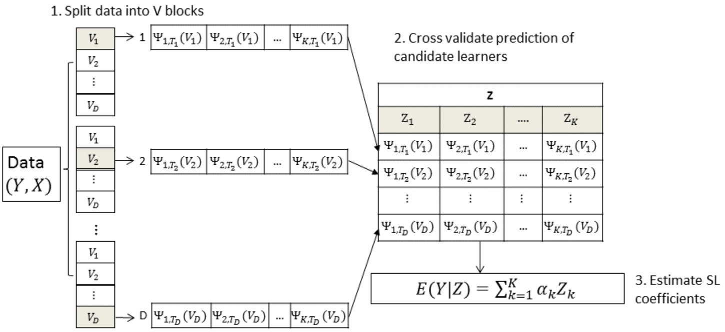
SL with Conditional Thresholding for Classification



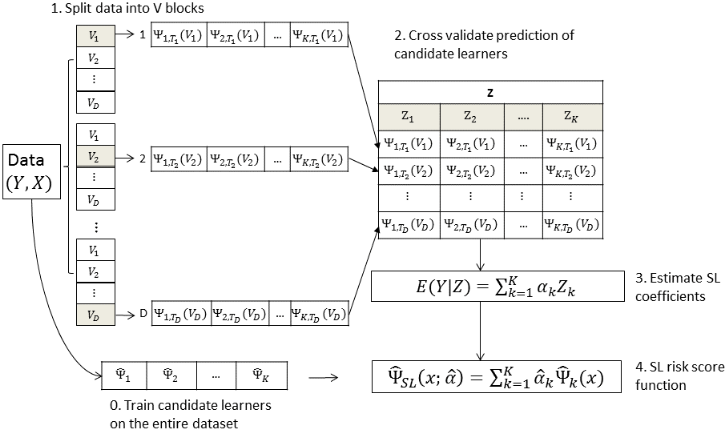
SL with Conditional Thresholding for Classification



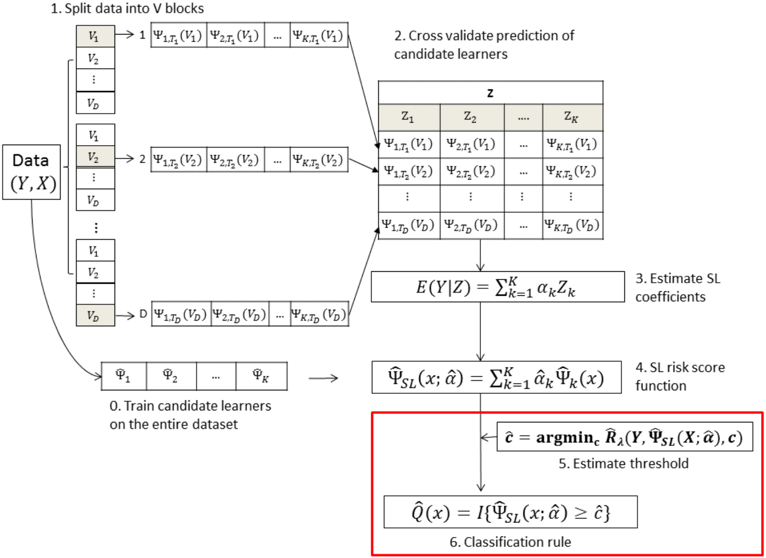
SL with Conditional Thresholding for Classification



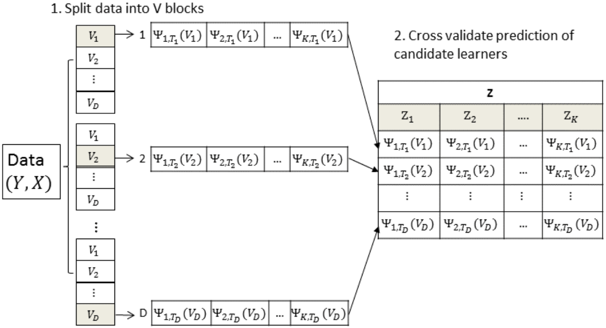
SL with Conditional Thresholding for Classification



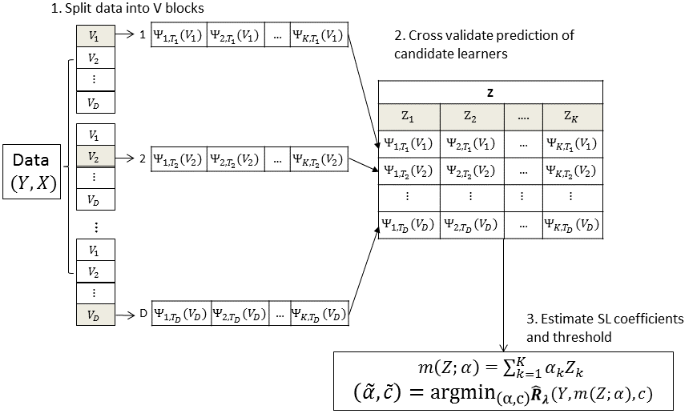
SL with Conditional Thresholding for Classification



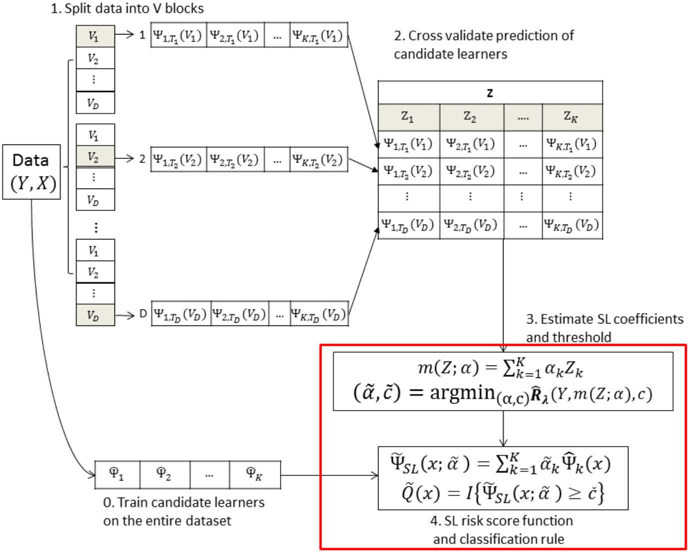
SL with Joint Thresholding for Classification



SL with Joint Thresholding for Classification



SL with Joint Thresholding for Classification



Empirical Weighted Misclassification Risk

$$\widehat{R}_\lambda(Y, \Psi(X;\alpha), c) = \frac{1}{n}\sum_{i=1}^{n} \lambda \mathbb{1}\{\Psi(X_i;\alpha) < c, Y_i = 1\}$$
$$+ (1-\lambda)\mathbb{1}\{\Psi(X_i;\alpha) \geq c, Y_i = 0\}$$

- Optimizing counts: computationally a very difficult problem
  - Lack of smoothness and convexity
  - Numerous optima

Minimization of WMR

- Some existing methods:
  - Approximate the WML with smooth solvable loss function: integrals of beta distribution to approximate the indicator functions in WML (Buja et al, 2005)
  - Hierarchical mathematical programming: linear program with equilibrium constraints (Mangasarian, 1994) for total misclassification loss
  - Hybrid accelerating algorithms: convex surrogate $max(1 + x, 0)$ of indicator function $\mathbb{1}(x > 0)$ (Chen and Mangasarian, 1996)

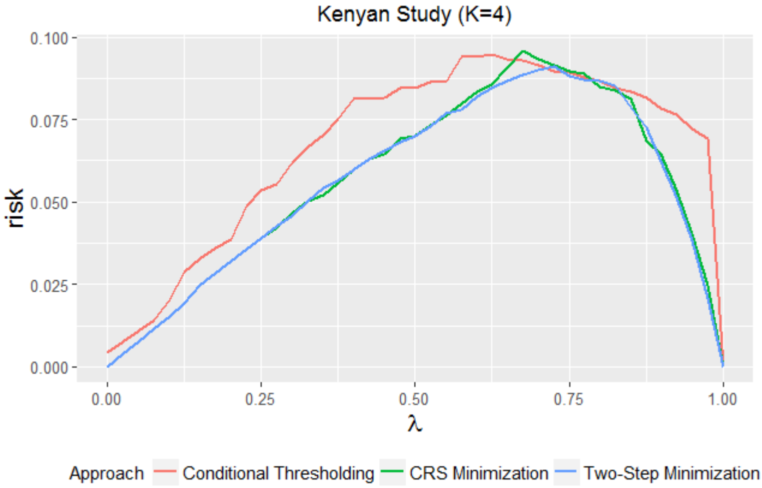
Minimization of WMR

- Our strategy
  - Direct search methods for global optimization
    - Controlled random search (Kaelo and Dixon, 2006)
    - Key: transform the problem into bounded region optimization
  - Two-step methods
    - Key: use a convex and continuous surrogate loss for estimating $\tilde{\alpha}$
    - Estimate $\tilde{c}$ based on $\tilde{\alpha}$
    - Can be extended to iterative procedures when the surrogate loss contains c

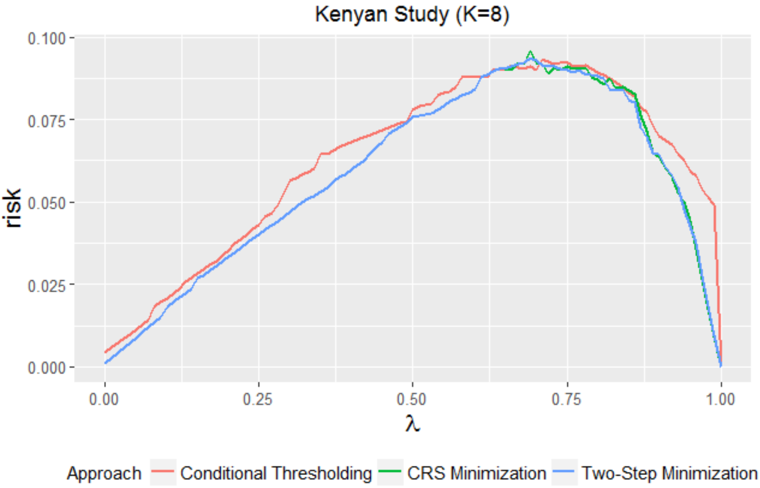
Simulations

- Settings:
  - Observe the variables used in outcome generation
  - Observe highly nonlinear transformations of the variables used in outcome generation
- Joint thresholding obviously outperforms conditional thresholding in the second setting; the two methods do not differ much in the first setting
- More candidate learners decrease the discrepancy
- Two-step and controlled random search have similar results; this have implications for computing

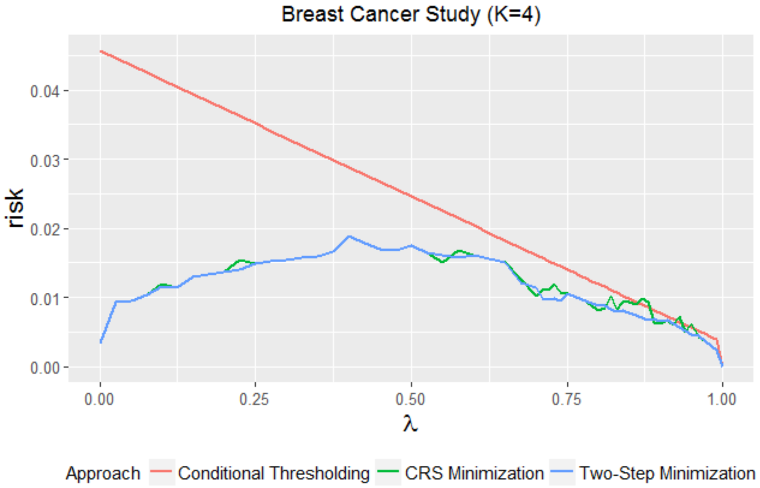
CV Weighted Misclassification Risk Stratified by SL Library Size K



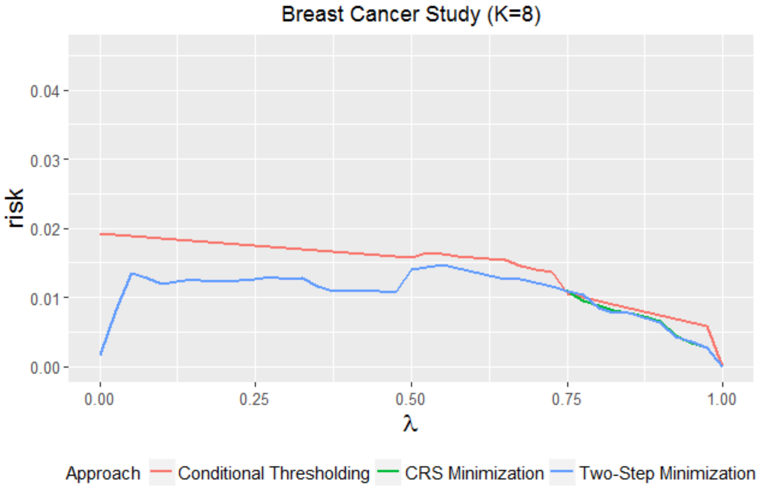
CV Weighted Misclassification Risk Stratified by SL Library Size K



CV Weighted Misclassification Risk Stratified by SL Library Size *K*



CV Weighted Misclassification Risk Stratified by SL Library Size K



Results

Kenyan HIV data
CT: conditional thresholding
CRS: joint thresholding using controlled random search

	λ = .2	
	CT	CRS
$\hat{\alpha}_{\text{random forest}}$	0.11	0.11
$\hat{\alpha}_{\text{logistic regression}}$	0	0
$\hat{\alpha}_{\text{quadratic splines}}$	0.42	0.42
$\hat{\alpha}_{\text{CART}}$	0	0
$\hat{\alpha}_{\text{10-NN}}$	0.20	0.20
$\hat{\alpha}_{\text{generalized boosting}}$	0.27	0.27
$\hat{\alpha}_{\text{SVM}}$	0	0
$\hat{\alpha}_{\text{Bagging}}$	0	0
$\hat{c}$	0.62	0.73

Results

	λ = .8	
	CT	CRS
$\hat{\alpha}_{\text{random forest}}$	0.11	0.04
$\hat{\alpha}_{\text{logistic regression}}$	0	0.19
$\hat{\alpha}_{\text{quadratic splines}}$	0.42	0.16
$\hat{\alpha}_{\text{CART}}$	0	0.33
$\hat{\alpha}_{\text{10-NN}}$	0.20	0.01
$\hat{\alpha}_{\text{generalized boosting}}$	0.27	0.06
$\hat{\alpha}_{\text{SVM}}$	0	0.12
$\hat{\alpha}_{\text{Bagging}}$	0	0.08
$\hat{c}$	0.16	0.18

Discussions

- Our work provides a general framework for using ensemble learners for binary classification and has the potential to be extended to more general threshold-based classification
- Joint thresholding performs as well as or better than the conditional thresholding approach in terms of properly estimating CV weighted misclassification risks
- In our analysis, difference between thresholding methods is smaller for larger SL library
- From Bayes' rule, optimal threshold at $1 - \lambda$ when $\Psi(X) = P(Y = 1|X)$. Threshold estimation is still very important!
- We anticipate the performance of our method to be comparable to threshold estimation based on CV SL predictions

Thank you!

Acknowledgment

- **Joseph Hogan (Advisor)**
- **Tao Liu (Co-advisor)**
- Rami Kantor
- Michael Daniels
- Allison Delong

Funding: NIH grants R01-AI-108441, P30-AI-42853

References

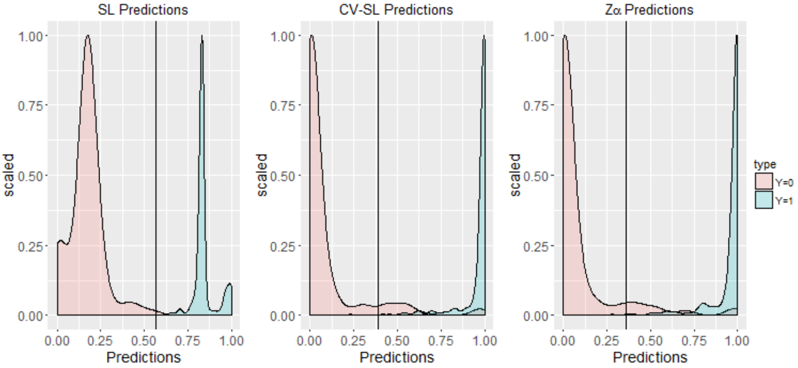
- A.W. van der Vaart, S. Dudoit, M.J. van der Laan. *"Oracle Inequalities for Multi-fold Cross Validation" (2006)*.
- Asuncion, Arthur, and David Newman. *"UCI Machine Learning Repository" (2007)*.
- K. Brooks, L. Diero, A. DeLong, M. Balamane, M. Reitsma, E. Kemboi, M. Orido, W. Emonyi, M. Coetzer, J. Hogan and others *"Treatment failure and drug resistance in HIV-positive patients on Tenofovir-based first-line antiretroviral therapy in western Kenya" (2016)*.
- A. Buja, W. Stuetzle, Y. Shen. *"Loss Functions for Binary Class Probability Estimation and Classification: Structure and Applications" (2005)*.
- C. Chen, O.L. Mangasarian. *"Hybrid Misclassification Minimization" (1996)*.
- L. Diero, A. DeLong, L. Schreier, E. Kemboi, M. Orido, M. Rono. *"High HIV Resistance and Mutation Accrual at low Viral Loads upon second Line Failure in western Kenya" (2014)*.

References

- M. Mann, L. Diero, E. Kemboi, F. Mambo, M. Rono, W. Injera, A. Delong, L. Schreier, K.W. Kaloustian, J. Sidle and others. *"Antiretroviral treatment interruptions induced by the Kenyan postelection crisis are associated with virological failure" (2013)*.
- E.C. Polley, M.J. van der Laan. *"Super Learner" (2007)*.
- E.C. Polley, M.J. van der Laan. *"Super Learner in Prediction" (2010)*.
- O.L. Mangasarian. *"Misclassification Minimization" (1994)*.
- P. Kaelo, M.M. Dixon. *"Some Variants of the Controlled Random Searchh Algorithm for Global Optimization" (2006)*.

Density Curves of SL Prediction, 10 fold CV-SL, $m(Z;\tilde{\alpha})$ on BRCA Data

Thresholds Estimated at $\lambda = 0.8$ ($K = 8$)



Contact

phone 203 824 0395
email yizhen_xu@brown.edu