A Novel Prediction and Classification Model for Medical Diagnosis

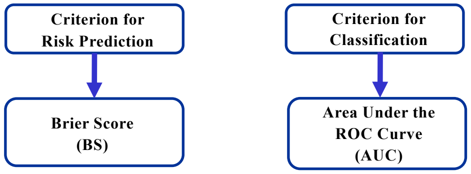
Minmin Zhang,[1] Zheng Zhang, Ph.D.[1]; Brian T. Denton, Ph.D.[1,2]

[1]Department of Industrial and Operations Engineering, University of Michigan, [2]Department of Urology, University of Michigan Medical School

Background

- Data prevail in the healthcare area, motivating the implementation of machine learning methods to improve the practice of medical diagnosis.
- Medical diagnosis involves two important decision problems: 1) whether the patients have a disease or not (classification) and 2) how likely the patients have a disease (risk prediction).
- How best to address these problems is a difficult question because of the conflicting criteria between classification and prediction.

Conflicting Criteria



Criterion for Risk Prediction

Brier Score (BS)

Criterion for Classification

Area Under the ROC Curve (AUC)

- Both risk prediction and classification properties are important, while the logistic regression gets one optimized in many cases.
- The goal of this research is to jointly optimize risk prediction and classification to trade off between the BS and the AUC.

Research Questions

- How to solve this multi-criteria problem efficiently for a near-to-optimal solution?
- What is the value of joint optimization compared with the standard approach of logistic regression?

Related Work

- AUC optimization vs. error rate minimization[1]
- Joint optimization of classification and linear regression[2]
- There are many papers about maximizing AUC, but the joint optimization of BS and AUC has not been studied yet.

Model

	Data with Positive Outcome	Date with Negative Outcome
Number of Data	I	J
Vector of Features	$X_i \in R^n$	$X_j \in R^n$
Predicted Probability	$Y_i \in [0,1]$	$Y_j \in [0,1]$

There are n features for each data set.

The decision variable is $\beta \in R^n$ such that β_k is the coefficient of feature k. We predict the probability based on the logit function as follows:

$$Y = \frac{1}{1+e^{-\beta X}}$$

Accuracy of Prediction

- Brier score (BS) measures the accuracy of probabilistic risk predictions based on the mean squared difference between the predicted probability and the actual outcome. Specifically,

$$BS = \frac{1}{I+J}\left(\sum_i (1-Y_i)^2 + \sum_j Y_j^2\right)$$

Accuracy of Classification

- AUC (area under the ROC curve) measures the accuracy of classification that is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (assuming 'positive' ranks higher than 'negative').
- An alternative expression of AUC is Wilcoxon-Mann-Whitney (WMW) statistic.

$$WMW = \frac{1}{IJ}\left(\sum_i \sum_j I(\beta X_i > \beta X_j)\right)$$

- In our model, we use a proxy function for WMW that is also an approximation of AUC.

$$\frac{1}{IJ}\left(\sum_i \sum_j S(\beta X_i\,, \beta X_j)\right)$$

where

$$S(\beta X_i\,, \beta X_j) = \frac{1}{1+e^{r(\beta X_i - \beta X_j)}}$$

in which r is a large constant number. Hence, we expect a low value of S function when positive and negative data are separated with βX_i greater than βX_j.

Cost function

- The goal is to minimize the following function:

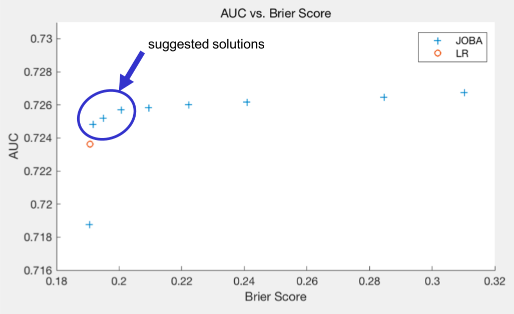
$$\frac{\lambda}{I+J}\left(\sum_i (1-Y_i)^2 + \sum_j Y_j^2\right) + \frac{1-\lambda}{IJ}\left(\sum_i \sum_j S(\beta X_i\,, \beta X_j)\right)$$

where λ is a weight factor to trade off between classification and risk prediction. The cost function is continuous and differentiable and therefore can be optimized over β using gradient descent.

Results

Test case: predicting and deciding the cellular localization sites of proteins (data source: UCI Machine Learning Repository[3]).

The graph shows how BS and AUC change when different weights are delegated. A set of Pareto solutions (JOBA) were solved by varying the weight factor from 0 to 1. The results are compared with the logistic regression model.



The optimized solutions demonstrate a trade off between classification and risk prediction. We suggested solutions that can perform simultaneously well in both classification and prediction.

Logistic regression solution is very close to the front line of the Pareto solutions, resulting in high-quality prediction but lower quality classification.

Conclusions

- It is possible to effectively solve the joint optimization of classification and risk prediction problem.
- Our model provides more flexible solutions that improve the trade off in important criteria for medical diagnosis compare to standard logistic regression.

Acknowledgements

This work was supported by the National Science Foundation (CMMI 0844511 to BTD); any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Reference

- [1] Cortes, Corinna, and Mehryar Mohri. "AUC optimization vs. error rate minimization." *Advances in neural information processing systems*. 2004.
- [2] Bertsimas, Dimitris, and Romy Shioda. "Classification and regression via integer optimization." *Operations Research* 55.2 (2007): 252-271.
- [3] "Expert Sytem for Predicting Protein Localization Sites in Gram-Negative Bacteria", Kenta Nakai & Minoru Kanehisa, *PROTEINS: Structure, Function, and Genetics* 11:95-110, 1991.