Selection and Combination of Markers for Prediction

NACC Data and Methods Meeting
September, 2010

Baojiang Chen, PhD
Sarah Monsell, MS
Xiao-Hua Andrew Zhou, PhD

Overview

1. Research motivation
2. Describe traditional methods
3. Discuss limitations
4. Describe smoothing & penalized methods
5. Illustrate proposed methods using NACC data
6. Discuss importance for researchers

Background

- Several recent papers on status of biomarker and early detection research:

  Daviglus ML, Bell CC, et al. NIH State-of-the-Science Conference Statement: Preventing Alzheimer's Disease and Cognitive Decline. *Ann Intern Med.* 2010;153:176-181.

  De Meyer G, Shapiro F, et al. Diagnosis-independent Alzheimer disease biomarker signature in cognitively normal elderly people. *Arch Neurol*. 2010;Aug 67(8):949-56.

- Wide variety in methodology
- Room for methodological development

Research Goals

Employ methodology that can:

1. Combine markers for prediction
2. Select a subset of candidate markers for efficiency

Note: Markers can be biomarkers, neuropsychological assessments, modifiable risk factors, etc.

Methods

Commonly used:

- Logistic regression
- Decision tree

Newer methods:

- PSAUC (penalized smoother area under the ROC curve)
- Smoothed area under the curve (AUC)-based approach

Logistic regression

- Logit link function from binomial distribution
- Interpretation for continuous markers:
  - β-coefficients are the additive effect on the log odds for a unit change in the marker
  - example: the log odds of dementia increases by β for each one unit increase in the marker
- Major assumption: logit link describes association between predictors & outcome

Decision tree

- Recursively partition the data into subsets that differ on risk
- Provide a tree-based algorithm for assigning risk scores

Limitations

Logistic regression & Classification Trees:

1. Do not penalize for inclusion of markers weakly associated with outcome

2. Rely on classification error to select markers

AUC vs classification error

- Traditional selection procedures (AIC, BIC, C_p) rely on classification error (accuracy) to select the best markers for prediction
- Classification error is determined from the number of subjects whose disease status was incorrectly classified by the prediction equation
- Better to use a ranking of patients in terms of likelihood of having the disease
- Ling et al (2003) showed that the area under the ROC curve (AUC) is a statistically consistent and more discriminating measure than classification error

Advantages of using AUC

- McIntosh et al. (2002) showed that, asymptotically, the AUC objective function yields the optimal linear combination of predictors in the sense that the corresponding ROC curve is optimal under the generalized linear model with an unknown link function: $P(D=1|\ Y)=g(X'\beta\ )$
- AUC(β): a combination of markers $X'\beta$

Empirical AUC-based method

$$AUC(\beta) = \frac{1}{n(n-1)} \sum_{D_i=1, D_j=0} I_{[X_i'\beta > X_j'\beta]}$$

Estimate AUC(β) by an empirical AUC:

- Find β corresponding to largest empirical AUC
- Relies on an indicator function
- Requires brutal search and development of special algorithms

Smoothing AUC-based methods

- Tackle computational difficulty by using a continuous function to approximate an indicator function
- Two ways for approximation:

  (a) a sigmoid approximation

  (b) a Probit approximation
- Resulting AUC is denoted by $S_A(\beta)$

Marker selection

- Although number of markers may be large, only a small number of markers may actually be associated with the binary clinical outcome
- Want to efficiently select the subset of significant markers
- There exists extensive literature on variable selection

Marker selection

- Traditional: stepwise and best subset selection procedures are
  - computationally intensive
  - hard to derive sampling properties
  - unstable
- Penalized estimation methods try to avoid these problems by shrinking estimates of regression coefficients toward zero relative to maximum likelihood (ML) estimates

Penalized AUC methods

- Apply the penalized variable selection method to the smoothed AUC function
- Can get different penalized smoothed AUC (PSAUC) functions:

$$L_n(\beta) = S_A(\beta) - p_{\lambda_n}(|\beta|)$$

smoothed AUC function

penalty function

PAUC

- The smoothly clipped absolute deviation (SCAD) penalty:
  - retains the good features of both subset selection and ridge regression
  - produces sparse solutions (many estimated coefficients are zero)
  - ensures continuity of the selected models (for the stability of model selection)
  - has unbiased estimates for large coefficients
- Lin et al (2010) and Zhou et al (2010) developed the PAUC with the SCAD penalty.

TGDR method

- Ma and Huang (2007) chose the gradient directed regularization method (TGDR)
- Sigmoid approximation to optimize AUC step-function

Summary of methods

Method	**Description**	**Marker & combination selection performed simultaneously**
Logistic regression	Logit link function	NO
Decision tree	Recursive partitioning of sample space, model each partition separately	SOME
PSAUC	Local smoothing of AUC curve Penalized for number of parameters Link function not specified Maximizes AUC and AUC penalty	YES
TGDR	Sigmoid approximation to optimize AUC step-function	YES

Advantages and limitations

Method	Advantages	Disadvantages
Logistic regression	Easy to implement, easy to interpret	Strong assumptions of link function, relies on classification error to select markers
Decision tree	Do not impose restrictions on functional form of associations, easy to implement & interpret	Forced to discretize inherently continuous markers, relies on classification error
PSAUC	Fewer model assumptions, more parsimonious models, less computationally burdensome than empirical AUC	Need to estimate an additional unknown parameter
TGDR	Fewer model assumptions, less computationally burdensome than empirical AUC	Need to estimate two additional unknown parameters

Relative performance

Simulation study performed by Lin et al (2010) found that:

- When the sample size was small, the logistic-based models had larger biases and variances.
- When the link function is true, the logistic regression is more efficient than PSAUC and TGDR. But, misspecification of the link function can lead to seriously biased estimators.
- When the sample size was small or medium, the PSAUC estimators had comparable variances with those of the logistic-based models but lower biases
- When the sample size was large, PSAUC and TGDR methods performed similarly

Cross-validation

- Used to estimate accuracy of the predictive model
- Evaluate how well the model will work with an independent data set
- Steps for k-fold cross-validation:

  1. partition data into k subsets
  2. k-1 subsets are used in model selection (training data)
  3. model is run on the remaining set (test data) and performance is recorded
  4. perform steps 2 & 3 k times & record the prediction accuracy of each subset
  5. average the results

Illustration with NACC data

- UDS does not have biomarker data so we rely on markers from MMSE & neuropsychology battery
- Goal: combine cognitive tests to diagnose progression from MCI to dementia within 2 years
- Sample :
  - UDS subjects
  - Diagnosed with MCI at baseline
  - No missing data on any of the cognitive tests
  - At least 2 years of follow-up
  - Total of 366 subjects

Variables for modeling

- Outcome: progression to dementia or not within two years of entry into the study (binary)
- Predictors: MMSE total score
  Logical Memory: Immediate & Delayed
  Digit Span: Forward & Backward
  Trail Making A & B
  WAIS Digit Symbol
  Boston Naming
  Category Fluency: Animals & Vegetables

Demographics of NACC sample

Variable	Mean /proportion	Stand. Dev.
Female	0.52	0.50
Age	72.71	8.99
Education	15.17	6.28
Demented	0.28	0.45

- Slightly more women than men
- More than half achieved some college education
- 28% progressed to dementia

Results

Markers considered for selection	Logit	Decision tree	PSAUC	TGDR
MMSE total	0.077	x	0	0
Current logical memory recall number	0.075	x	0	0
Digit span forward trials correct	0.026	x	0	0
Digit span forward length	0.372		0	0
Digit span backward trials correct	-0.627		-0.995	0
Digit span backward length	0.575		0	0
Total animals named	-0.075		0	0
Total vegetables named	-0.315	x	0	0
Trail A seconds to completion	0.002	x	0	0
Trail A number commission errors	-0.088		0	0
Trail A number correct lines	0.018		0	0
Trail B seconds to completion	0.005	x	0.105	1.000
Trail B number commission errors	-0.024		0	0
Trail B number correct lines	-0.033		0	0
WAIS total correct items	-0.041	x	0	0
Delayed logical memory recall number	-0.097		0	0
Total score Boston naming test	-0.027	x	0	0
Run time (minutes)	0:01	0:02	2:00	2:30
AUC (from 5-fold cross validation)	0.600	0.627	0.685	0.645

Discussion of results

- PSAUC method has highest AUC
- Final prediction model based on PSAUC:

g(P(progression to dementia)) = -0.995*Digit span backward trials correct + 0.105*Trail B seconds to completion

where g() is some unknown function

- Most methods result in similar AUCs and choose similar predictors (but will not always be the case)
- Use of NACC data is for illustration of methods only- sample does not represent any real target population

Limitations

- Excluded significant amount of data from analysis due to missing values on cognitive tests
- New methods better suited for studies with more variables and fewer subjects (i.e. pixel-based image data)
- New methods take much longer to run
- No available software packages for new methods

What does this mean for researchers?

- When to consider new methods:
  - When the goal is to identify a combination of predictors & not necessarily interpret parameter estimates (AUC-based methods)
  - When the number of candidate markers is large relative to the sample size (penalized methods such as PSAUC)
- New methods often outperform conventional methods in these contexts

References

- Daviglus ML, Bell CC, et al. NIH State-of-the-Science Conference Statement: Preventing Alzheimer's Disease and Cognitive Decline. *Ann Intern Med.* 2010;153:176-181.
- De Meyer G, Shapiro F, et al. Diagnosis-independent Alzheimer disease biomarker signature in cognitively normal elderly people. *Arch Neurol*. 2010; Aug 67(8):949-56.
- McIntosh MW and Pepe MS. Combining several screening tests: optimality of the risk score. Biometrics. 2002; 58: 657-664
- Lin H, Zhou L, Peng H, Zhou XH. Selection and combination of biomarkers using ROC method for disease classification and prediction. *Canadian Journal of Statistics*. 2010; in press.
- Ling C. X., Huang, J., and Zhang H. AUC: a statistically consistent and more discriminating measure than accuracy. In Proceedings of 16th Canadian Conference on Artificial Intelligence. 2003.
- Ma S, Huang J. Combining multiple markers for classification using ROC. *Biometrics*. 2007; 63: 751-757.
- Zhou XH, Chen B, Xie YM, Tian F., Liu H, and Liang X. Variable selection using the optimal ROC curve: An application to a Traditional Chinese Medicine study on osteoporosis disease. *Statistics in Medicine 2010*. In press.

Questions?

Thank the NACC methods group for offering advice and support for this project