Maximization of the Partial Area under the ROC curve using a Boosting Technique

Osamu Komori, Prediction and Knowledge Discovery Research Center
komori@stat.math.ntu.edu.tw

Introduction

- This research attempts to improve the accuracy of discrimination of the diseased group from non-diseased group. The accuracy is often measured by the partial area under the ROC curve (pAUC) [5].
- The association between markers and outcome variable (disease or non-disease) is also important in medical or biological sciences.
- We have developed a new statistical method that aims to maximize the pAUC using a boosting technique.
- The resultant score plots show the association in a visually-apparent way.

Status quo

- does not take into account the non-linear structure of the association between markers and outcome variable[6][7].
- Moreover, Eguchi and Copas [1] and McIntosh and Pepe [4] show the theorem about pAUC: Age, Size, Grade, Angi, ERp, PRp and Lymp

ROC curve and pAUC

\[ \Phi(\gamma) = \text{pAUC}(F + \gamma m(X, \alpha, \alpha)) \]

The pAUC is more suitable for clinical setting in which a high true positive rate is required with a low false positive rate.

Theorem about pAUC

Theorem 1. For a pair of fixed \( \alpha_1 \) and \( \alpha_2 \), let

\[ \Phi(\gamma) = \text{pAUC}(F + \gamma m(X, \alpha_1, \alpha_2)) \]

\( \gamma \) is a scalar, \( m(X) = g_1(x)/g_2(x) \) and \( m(X) \) is a strictly increasing function of \( \gamma \) and \( \lim_{\gamma \to \infty} \text{pAUC}(F, \alpha_1, \alpha_2) = \lim_{\gamma \to -\infty} \Phi(\gamma) = \text{pAUC}(\lambda, \alpha_1, \alpha_2) \)

As seen in Theorem 1, the approximate pAUC has no maximum, but a supremum. Hence, we will consider the penalty term in the objective function in order to ensure the existence of the maximum and make the pAUCBoost algorithm numerically stable.

Objective function

- Prepare a set of weak classifiers, from which we construct a score function \( F(x) \).

\[ F \left( \frac{f(x)}{m(X)} \right) = N_p, \quad (z_i \leq k) \]

The basis functions of the natural cubic spline for \( z_k \) are defined as

where \( d_k(x) = (z_k - \xi_{k-1})^2 - (z_k - \xi_{k+1})^2, \) \( \xi_{k-1} \) and \( \xi_{k+1} \) denotes the positive part of \( z \). The standardization factor \( Z \) for \( N_p \) is given as

\[ Z = \begin{cases} 1, \quad l = 1, \\ \xi_{l-1} - \xi_l, \quad l = 2, \\ N_{l}(\xi_{l-1}) - N_{l}(\xi_{l+1}), \text{ otherwise,} \end{cases} \]

\( \xi_0 = \xi_{m+1} \)

and \( \xi_{m+1} \) is one of \( m \) knots \( \xi_{k} < \xi_{k+1} < \cdots < \xi_{k+m} \) for \( z_k \).\n
- The objective function we propose is

\[ \text{pAUCBoost}(F, \pi_1, \pi_2) = \text{pAUC}(F, \pi_1, \pi_2) - \lambda \sum_{n=1}^{p} \left( f_n^2(z_k) \right) \left( \frac{d_n(x)}{x} \right) \]

\( \text{pAUC} \) is the second derivative of the \( k \)-th component of \( F(x) \), and \( \lambda \) is a smoothing parameter that controls the smoothness of \( F(x) \).

- Without loss of generality, we can rewrite it as

\[ \text{pAUCBoost}(F, \pi_1, \pi_2) = \text{pAUC}(F, \pi_1, \pi_2) - \lambda \sum_{n=1}^{p} \left( f_n^2(z_k) \right) \left( d_n(x)/x \right) \]

Note that the maximizer of the objective function is shown to be the natural cubic splines [5].

pAUCBoost algorithm

1. Start with a score function \( F(x) = 0 \) and set each coefficient \( \beta(k) \) of weak classifiers to be \( 1 \) or \( -1 \).
2. For \( l = 1, \ldots, \ldots, T \)
   a. Calculate the values of thresholds \( \pi_1, \pi_2 \) for each \( F_{l-1} = \beta(l) \cdot f(l) \).
   b. Update \( \beta(l) \) to \( \beta(l) \) with one-step Newton-Raphson iteration.
   c. Find the best weak classifier \( f(l) \)
   d. Update the score function as
   e. Finally, output a final score function \( F(x) = \sum_{n=1}^{p} f_n(x) \).

Breast cancer data

- Two types of data [7]
  - clinical data: Age, Size, Grade, Angi, ERp, PRp and Lymp
  - genomic data: gene expression profiles (20000 genes)
- training data: 78 patients; test data 19 patients
- We apply AUCBoost to clinical data using natural cubic splines to Age and Size (continuous markers), and decision stumps to the others (discrete markers)

Figure 1. (a) Score plots of clinical markers; (b) Clustering cited from [7].

- The resultant AUC is 0.892 and 0.899 for training and test data, resp.
- After the pAUC-based filtering process, we apply pAUCBoost with natural cubic splines to the 11 genes.

Figure 2. Score plots of the selected 11 genes.

Figure 3. (a) The pAUC for training data (black) and test data (gray); (b) the results of van’t Veer [7] .

- The resultant pAUCs for both training and test data are more than 3 times bigger than their results: 0.025 and 0.0008, resp. [7].

References