

MAXIMIZING AREA UNDER THE RECEIVER OPERATING CHARACTERISTIC CURVE FOR BIOMARKER COMBINATION

Yuxuan Chen* and Yijian Huang

Emory University

Abstract: Multiple biomarkers are often combined for more accurate disease diagnosis. For this purpose, one popular performance metric is the area under the receiver operating characteristic (ROC) curve (AUC). Optimizing the empirical AUC over linear combinations of biomarkers, however, faces two primary challenges. First, AUC is scale-invariant to the linear combinations, creating difficulties in both the computation and asymptotic study. Most available approaches actually consider a restricted problem by setting one coefficient to a constant. Second, the empirical AUC is piecewise-constant and standard gradient-based computational algorithms are not applicable. Existing methods maximize kernel-smoothed AUC instead, but they can be sensitive to bandwidth choice. In this article, we tackle these challenges by developing a new empirical AUC maximization method. Computationally efficient algorithms are provided for both the point and variance estimation of the estimated combination coefficients. Simulation studies show good computational and statistical performance of the proposed methods. An illustration is provided with a clinical application.

Key words and phrases: Classification, kernel smoothing, optimization algorithm, receiver operating characteristic curve, variance estimation.

1. Introduction

With advances in high-throughput sequencing and molecular profiling technologies, many biomarkers have been identified for the purpose of disease diagnosis. As a single biomarker may not be sufficiently informative, combining multiple biomarkers holds the promise for improved accuracy. For this purpose, likelihood-based methods such as logistic regression are widely adopted. Despite their good computational properties and wide acceptance, such a combination may be suboptimal in the case of model misspecification. As an alternative and more robust method, Pepe, Cai and Longton (2006) considered the area under the receiver operating characteristic curve (AUC) and suggested combining biomarkers to maximize the empirical AUC. AUC is a popular performance metric and can be interpreted as the probability that a diseased individual has a larger biomarker combination than a healthy individual.

*Corresponding author. E-mail: cyxolivia@gmail.com