Boosting Algorithm that Maximizes the Area Under ROC Curve

張源俊

中央研究院統計科學研究所

Abstract

Boosting is one of the most successful ensemble classifiers and has attracted much attention recently. At each stage of boosting algorithm will train a weak base learner (classifier) using the re-weighted training sample, which are based on the performance of the previous weak base learner (classifier). Then the final classifier is an ensemble of the sequence of weak base learners obtained in the boosting procedure. There are both theoretical and empirical results show that the boosting procedure can actually improve the performance of some weak based learners in terms of prediction accuracy.

From statisticians' point of view, the prediction accuracy is not the only assessment measure for evaluating classifiers. On the other hand, the Receiver Operation Characteristic (ROC) curve is a basic assessment tool in biostatistics and psychometrics. The related literature of ROC curve may date back to 1950's. Recently, ROC curve is getting popular in machine learning.

Since Shapiro (1990) raised the idea of boosting, there are many modifications of it published in machine learning and Statistics literature. In this report, we are going to integrate the ROC curve criterion into boosting algorithms, such that the final ensemble of weak learners will maximize the area under ROC curve. The technique we used is similar to that of combining many bio-makers in bio-informatics classification/diagnostic problems. For illustration purpose, we will start from a binary classification with the independent variables following a multivariate normal distribution. The original Adaboost algorithm is used for demonstration only.