ARTICLE IN PRESS

Computational Statistics and Data Analysis ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Nonparametric additive model with grouped lasso and maximizing area under the ROC curve

Sungwoo Choi, Junyong Park *

Department of Mathematics and Statistics, University of Maryland Baltimore County, 21250 MD, USA

ARTICLE INFO

Article history:
Received 6 September 2013
Received in revised form 11 February 2014
Accepted 21 March 2014
Available online xxxx

Keywords:
ROC curve
AUC
Additive model
Variable selection
High dimension

ABSTRACT

An ROC (Receiver Operating Characteristic) curve is a popular tool in the classification of two populations. The nonparametric additive model is used to construct a classifier which is estimated by maximizing the U-statistic type of empirical AUC (Area Under Curve). In particular, the sparsity situation is considered in the sense that only a small number of variables is significant in the classification, so it is demanded that lots of noisy variables will be removed. Some theoretical result on the necessity of variable selection under the sparsity condition is provided since the AUC of the classifier from maximization of empirical AUC is not guaranteed to be optimal. To select significant variables in the classification, the grouped lasso which has been widely used when groups of parameters need to be either selected or discarded simultaneously is used. In addition, the performance of the proposed method is evaluated by numerical studies including simulation and real data examples compared with other existing approaches.

1. Introduction

In binary classification problems, it is one possible way to evaluate the performance of classifiers based on the receiver operating characteristic (ROC) curve. The ROC curve has been originally introduced in signal detection theory, and it became one of the important tools to measure the performance of classification rules. The advantage of the ROC curve is the fact that it does not assume the parametric form of class distributions and it is independent of underlying misclassification costs (Provost and Fawcett, 1997; Pepe et al., 2006). An ROC curve is the plot of true positive rate (TPR) on the horizontal axis and false positive rate (FPR) on the vertical axis as the classification threshold varies. In the ROC space, the upper left corner represents perfect classification while a diagonal represents random allocation of individuals to one of the two classes. In practice, the ROC curve is a single curve lying between these two extreme cases, so it is plotted in the upper triangle of the graph. For example, in Fig. 1.1, the closer it lies to the upper left corner, the better is the performance of the classifier to discriminate between two groups.

Even though the ROC curve provides a useful summary of information about a description of the performance of a classifier, it is often preferable to employ a single scalar value such as the area under the ROC curve (AUC) for the comparison of different classifiers. For instance, a larger AUC value indicates on average a better classifier performance, even though it is possible that a classifier with a higher AUC can be outperformed by a lower AUC classifier at some region of the ROC space. It has been established that AUC is statistically consistent and more discriminating than predictive accuracy (Ling et al.,

* Corresponding author. Tel.: +1 410 455 2407.
E-mail address: junpark@umbc.edu (J. Park).

http://dx.doi.org/10.1016/j.csda.2014.03.010
0167-9473/

Please cite this article in press as: Choi, S., Park, J., Nonparametric additive model with grouped lasso and maximizing area under the ROC curve. Computational Statistics and Data Analysis (2014), http://dx.doi.org/10.1016/j.csda.2014.03.010

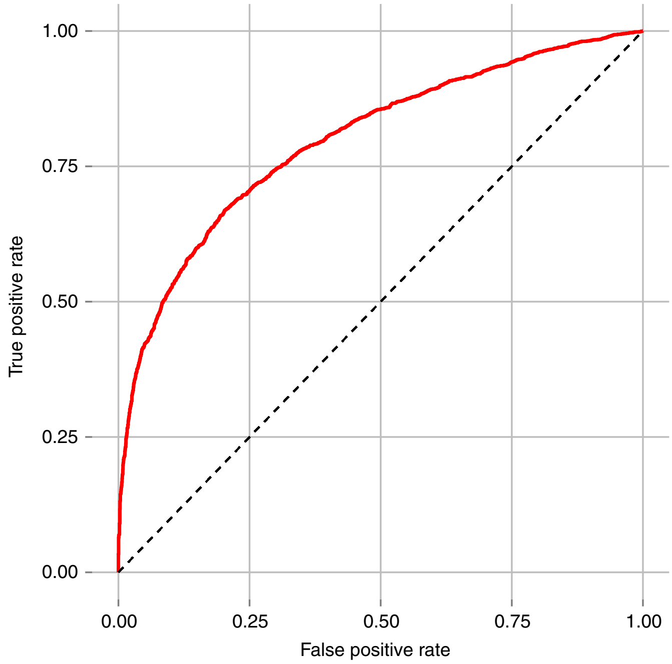

Fig. 1.1. An ROC curve.

2003). Moreover, AUC is a suitable measure to evaluate the ability of classifiers to rank instances in two-class classification problems (Ataman et al., 2006).

When the multiple measurements are taken on each subject, it is a common practice to consider the classification rule based on a single score S which is some continuous function of the predictors. The AUC of a classifier expresses the probability that a randomly selected individual from one group gets a higher score by the classifier than a randomly selected individual from the other group. More specifically, when $\mathbf{X}$ and $\mathbf{Y}$ are random vectors of measurements from two groups, AUC can be expressed as $P(S(\mathbf{X}) > S(\mathbf{Y}))$. Since the optimal ROC curves have the maximum AUC, Pepe et al. (2006) proposed to use the empirical AUC as an objective function for combining multiple predictors, and simply illustrated that the AUC approach has better performances compared to the logistic regression approach. For high dimensional data, Ma and Huang (2005) considered the use of the smooth sigmoid function to approximate the indicator function in the empirical AUC. Instead of using the sigmoid function, Zhao et al. (2011) proposed surrogate loss functions to approximate the non-convex AUC.

In the multivariate context, the linear combinations of the predictors is the most widely used rule so that the combined score achieves the optimal AUC over all possible linear combinations. Even though representing the score by a traditional linear model is an easy and convenient approximation, it often fails in many practical situations which are often not linear. In this paper, we consider a more general case such as the nonparametric additive model which will be discussed in Section 3.

In particular, when there are lots of variables, it is commonly assumed that only a small fraction of variables are significant, namely sparsity. For example, such a sparsity assumption is used in the lasso by Tibshirani (1996), the least angle regression (LARS) by Efron et al. (2004) and the threshold gradient directed regularization method by Friedman and Popescu (2004). We present some asymptotic result claiming that the empirical risk minimization may not be guaranteed to obtain the optimal AUC when the number of variables is much larger than the sample size. However, when there are too many noisy variables, it is expected that we obtain much better AUC via the empirical minimization from selected variables considered as significant variables in the classification even if the dimension is much larger than sample sizes.

Two main approaches will be considered such as optimizing empirical AUC with the grouped lasso and logistic regression with the grouped lasso (Meier et al., 2008). In addition, we consider the element-wise sparsity to compare with the group sparsity in simulation studies. In both approaches, the situation of high dimensionality with lots of noisy variables will be considered, which demands variable selection to improve the AUC. It is expected that minimization of empirical AUC obtains larger AUC than logistic regression does since the empirical AUC is more appropriate as a loss function in maximizing AUC than the likelihood function in logistic regression. Also, considering group structure is expected to yield a better classification performance than the element-wise structure in higher order polynomial expansions of variables. We evaluate the performances of all approaches through numerical simulations and real data examples.

The objective of this paper is to extend the idea of maximizing empirical AUC as an objective function based on the linear model to the nonparametric additive model with appropriate basis functions. In particular, when we want to consider more flexible functions other than linear models and to deal with categorical or factor variables in a high dimensional setting, it is

Please cite this article in press as: Choi, S., Park, J., Nonparametric additive model with grouped lasso and maximizing area under the ROC curve. Computational Statistics and Data Analysis (2014), http://dx.doi.org/10.1016/j.csda.2014.03.010