Ann Inst Stat Math (2011) 63:961–979
DOI 10.1007/s10463-009-0264-y

A boosting method for maximization of the area under the ROC curve

Osamu Komori

Received: 1 December 2008 / Revised: 8 July 2009 / Published online: 28 October 2009

Abstract We discuss receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC) for binary classification problems in clinical fields. We propose a statistical method for combining multiple feature variables, based on a boosting algorithm for maximization of the AUC. In this iterative procedure, various simple classifiers that consist of the feature variables are combined flexibly into a single strong classifier. We consider a regularization to prevent overfitting to data in the algorithm using a penalty term for nonsmoothness. This regularization method not only improves the classification performance but also helps us to get a clearer understanding about how each feature variable is related to the binary outcome variable. We demonstrate the usefulness of score plots constructed componentwise by the boosting method. We describe two simulation studies and a real data analysis in order to illustrate the utility of our method.

Keywords AUC · Boosting · Classification · ROC curve · Smoothing

1 Introduction

The receiver operating characteristic (ROC) curve has been widely used in medical and biological sciences (Zhou et al. 2002; Pepe 2003), for applications in which the classification performance can be measured by the area under the ROC curve (AUC). This curve has three primary appealing properties. First, it does not assume any specific distributional model, so a method based on the ROC is distribution-free, in contrast to logistic regression analysis or classical linear discriminant analysis

O. Komori (✉)
Department of Statistical Science, The Graduate University for Advanced Studies,
Minami-azabu, Tokyo, 106-8569, Japan
e-mail: komori@ism.ac.jp

under normality assumption. Second, it is independent of the prior probabilities of group membership, so it is able to accommodate case–control studies. Third, the AUC is not influenced by the choice of thresholds that may be changed according to each decision-maker's objective; hence, the AUC expresses the intrinsic accuracy of classification performance. The advantages of the AUC over the odds ratio or relative risk when evaluating the classification performance are discussed by Pepe et al. (2004).

A procedure for maximizing the AUC using a linear combination of multiple feature variables has been proposed (Pepe and Thompson 2000) in order to improve on diagnostic accuracy of a single feature variable, and Pepe et al. (2006) have shown that the AUC-based method can be far superior to logistic regression in certain situations. Ma and Huang (2005) extended this strategy to high-dimensional data by adopting a sigmoid approximation for the AUC. The assumption of linearity gives us easily interpretable results of the analysis, and allows us to get the rough characteristics of each feature variable. However, this strict assumption is often unable to capture informative nonlinear structures in the real world.

Moreover, it has been proved that the optimal combination of feature variables that maximizes the AUC is constructed based on the likelihood ratio (Eguchi and Copas 2002; McIntosh and Pepe 2002). This implies that even under a simple setting such as a normality assumption with unequal covariance matrices, the optimal combination is not linear but quadratic. Further details are described in Sect. 4.2.

In this paper, we propose a new statistical method to detect a more essential association between feature variables and a binary outcome variable using a boosting technique, and apply the method to the combination of the feature variables for better classification. A typical one of the boosting methods is AdaBoost (Freund and Schapire 1997), which is designed to minimize the exponential loss. An AdaBoost-based boosting method for the AUC is presented by Long and Servedio (2007), along with its theoretical justification. The purpose of boosting methods is to construct a strong classifier by combining various weak classifiers. Recently, a variety of loss functions other than the exponential loss have been proposed and discussed in several contexts (Murata et al. 2004).

On the other hand, the generalized additive model (GAM) proposed by Hastie and Tibshirani (1986) has wide applications in a variety of research fields. This is mainly because this model can detect the nonlinear effects of feature variables on the objective function flexibly, without sacrificing interpretability:

$$\eta(E(y|\boldsymbol{x})) = F_1(x_1) + \cdots + F_p(x_p),$$

where $\boldsymbol{x} = (x_1, \ldots, x_p)'$, η is a link function and F_k, $k = 1, \ldots, p$, are unspecified functions of x_k. Thus, GAM is also well suited for binary classifications in medical and biological fields, in which the association of the feature vector $\boldsymbol{x}$ with an outcome variable y is of great interest. We consider a model, similar to GAM, that attaches importance to interpretability as well as flexibility, maximizing the AUC for a score function $F(\boldsymbol{x})$ by a boosting algorithm. As a result, we obtain $F(\boldsymbol{x})$ of the form

$$F(\boldsymbol{x}) = F_1(x_1) + \cdots + F_p(x_p),$$

in which we consider score plots of $F_k(x_k)$ against the kth feature variable x_k. These plots are useful in association studies, for looking at how each feature variable works in the classification and for detecting which feature variable is the most effective one.

This paper is organized as follows. In Sect. 2, we give a brief review of the ROC curve and discuss the relationship between the AUC and the approximate AUC. In Sect. 3, we propose AUCBoost, a new boosting method based on the maximization of the AUC. In Sect. 4 we present two simple simulation studies to investigate the efficiency of AUCBoost, and in Sect. 5 we demonstrate the application of AUCBoost to a real data set. We close Sect. 6 with concluding remarks and ideas for future work.

2 Receiver operating characteristic curve

2.1 AUC

Let y be a binary class label ($y = 0, 1$), $\boldsymbol{x} \in \mathbf{R}^p$ be a feature vector, and $g_0(\boldsymbol{x})$, $g_1(\boldsymbol{x})$ be probability density functions for each class. We classify a subject with feature vector $\boldsymbol{x}$ into class 1 if a score function $F(\boldsymbol{x})$ is greater than or equal to a threshold value c, and into class 0 otherwise. Then, the false positive rate (FPR) and true positive rate (TPR) are defined as

$$\mathrm{FPR}(c) = \int_{F(\boldsymbol{x}) \geq c} g_0(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \quad \text{and} \quad \mathrm{TPR}(c) = \int_{F(\boldsymbol{x}) \geq c} g_1(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \tag{1}$$

By pairing these probabilities, the ROC curve is given as

$$\mathrm{ROC} = \{(\mathrm{FPR}(c), \mathrm{TPR}(c)) \mid c \in \mathbf{R}\},$$

which is illustrated in Fig. 1.

From (1), the AUC is written as

$$\mathrm{AUC}(F) = \int_{\infty}^{-\infty} \mathrm{TPR}(c) \mathrm{dFPR}(c). \tag{2}$$

The large separation of $g_0(\boldsymbol{x})$ and $g_1(\boldsymbol{x})$ could make the AUC close to 1. However, note that it is also dependent on a score function $F(\boldsymbol{x})$, which we must determine in an analysis of data. Only after employing an adequate $F(\boldsymbol{x})$ for the two probability density functions can we obtain the best value of the AUC. Equation (2) can be expressed in another manner:

$$\mathrm{AUC}(F) = P(F(\boldsymbol{X}_1) \geq F(\boldsymbol{X}_0)),$$

where $\boldsymbol{X}_0$, $\boldsymbol{X}_1$ are independent p-dimensional random vectors from class 0 and class 1, respectively (Bamber 1975). The empirical AUC for given observations $\{\boldsymbol{x}_{0i} : i$

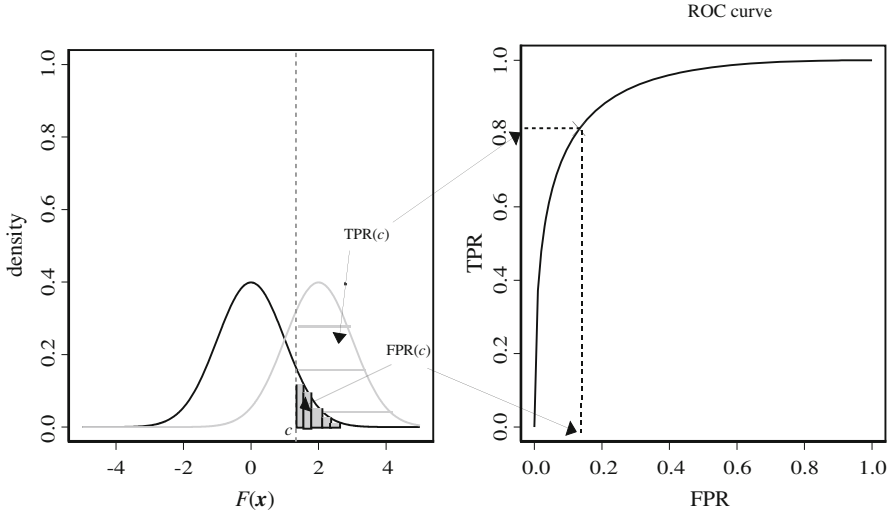

Fig. 1 The *left panel* illustrates the definition of FPR and TPR with two probability density functions of $F(\boldsymbol{x})$ for class 0 (*black*) and 1 (*gray*), and a threshold c. The *right panel* is the corresponding ROC curve

$= 1, \ldots, n_0\}$ of the class 0 and $\{\boldsymbol{x}_{1j} : j = 1, \ldots, n_1\}$ of the class 1 is given by

$$\overline{\mathrm{AUC}}(F) = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \mathrm{H}(F(\boldsymbol{x}_{1j}) - F(\boldsymbol{x}_{0i})), \tag{3}$$

where $\mathrm{H}(z)$ is the Heaviside function: $\mathrm{H}(z) = 1$ if $z \geq 0$ and 0 otherwise. In the case that $F(\boldsymbol{x})$ is discrete or there are tied values between $F(\boldsymbol{x}_{0i})$ and $F(\boldsymbol{x}_{1j})$, $\mathrm{H}(z)$ is replaced with $\mathrm{H}^*(z)$ that is defined to be 1 if $z > 0, \frac{1}{2}$ if $z = 0$ and 0 if $z < 0$.

2.2 Approximate AUC

We would like to obtain an optimal score function in the sense of maximizing the AUC in a class of score functions. It is known that the error rate is minimized by Bayes rule (McLachlan 2004), which can be expressed using a strictly increasing function of the likelihood ratio. Similarly, the Neyman–Pearson lemma (Neyman and Pearson 1933) establishes that the ROC curve for an arbitrary score function is everywhere below the ROC curve for the likelihood ratio (Eguchi and Copas 2002; McIntosh and Pepe 2002). That is, the optimal score function that maximizes the AUC is given as

$$F(\boldsymbol{x}) = m\left(\Lambda(\boldsymbol{x})\right), \tag{4}$$

where $\Lambda(\boldsymbol{x}) = g_1(\boldsymbol{x})/g_0(\boldsymbol{x})$ and m is a strictly increasing function. In this way, we observe that the maximization of the AUC is equivalent to the minimization of the error rate in the sense of Bayes rule.

In practice, the maximization of the empirical AUC presents some difficulties because it consists of a sum of nondifferentiable functions, as seen in Eq. (3). This feature prevents us from using gradient-based methods and requires a time-consuming search for the optimal score function (Pepe and Thompson 2000; Pepe et al. 2006). However, such a method becomes impossible to implement as the number of feature variables increases greatly. Therefore, as a means of maximizing the empirical AUC, it has become common to use smooth-function approximations. Eguchi and Copas (2002) used the standard normal distribution function, and Ma and Huang (2005) proposed a sigmoid approximation for this purpose. In this paper, we consider the former approximation:

$$\overline{\mathrm{AUC}}_\sigma(F) = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \mathrm{H}_\sigma(F(\boldsymbol{x}_{1j}) - F(\boldsymbol{x}_{0i})),$$

where $\mathrm{H}_\sigma(z) = \Phi(z/\sigma)$, with Φ being the standard normal distribution function. A smaller scale parameter σ means a better approximation of the Heaviside function $\mathrm{H}(z)$. The choice of the approximation function of $\mathrm{H}(z)$ does not matter so much; the important property is that the first derivative of the approximation function must be symmetric, which is satisfied in both $\mathrm{H}_\sigma(z)$ and the sigmoid function. This property is essential for the proof of Theorem 1.

Next, we discuss the relationship between the AUC and the approximate AUC. We note that the AUC for a score function $F(\boldsymbol{x})$ has an integral formula given as

$$\mathrm{AUC}(F) = \iint \mathrm{H}(F(\boldsymbol{x}_1) - F(\boldsymbol{x}_0)) g_0(\boldsymbol{x}_0) g_1(\boldsymbol{x}_1) \mathrm{d}\boldsymbol{x}_0 \mathrm{d}\boldsymbol{x}_1.$$

Similarly, the approximate AUC is given as

$$\mathrm{AUC}_\sigma(F) = \iint \mathrm{H}_\sigma(F(\boldsymbol{x}_1) - F(\boldsymbol{x}_0)) g_0(\boldsymbol{x}_0) g_1(\boldsymbol{x}_1) \mathrm{d}\boldsymbol{x}_0 \mathrm{d}\boldsymbol{x}_1.$$

Hence, we observe that $\overline{\mathrm{AUC}}_\sigma(F)$ almost surely converges to $\mathrm{AUC}_\sigma(F)$ as n_0 and n_1 both increase to infinity.

Theorem 1 *Let*

$$\Psi(c) = \mathrm{AUC}_\sigma\left(F + c\, m\left(\Lambda\right)\right),$$

where $\Lambda(\boldsymbol{x}) = g_1(\boldsymbol{x})/g_0(\boldsymbol{x})$ *and* m *is a strictly increasing function. Then,* $\Psi(c)$ *is a strictly increasing function of* $c \in \mathbf{R}$*, and*

$$\sup_F \mathrm{AUC}_\sigma(F) = \lim_{c \to \infty} \Psi(c) = \mathrm{AUC}\left(\Lambda\right). \tag{5}$$

Proof Let $\zeta(\boldsymbol{x}) = m(\Lambda(\boldsymbol{x}))$. Then, the first derivative of $\Psi(c)$ with respect to c is given as

$$\iint (\zeta(\boldsymbol{x}_1) - \zeta(\boldsymbol{x}_0))\, \mathrm{H}'_\sigma\left(F(\boldsymbol{x}_1) + c\,\zeta(\boldsymbol{x}_1) - F(\boldsymbol{x}_0) - c\,\zeta(\boldsymbol{x}_0)\right) g_0(\boldsymbol{x}_0) g_1(\boldsymbol{x}_1) \mathrm{d}\boldsymbol{x}_0 \mathrm{d}\boldsymbol{x}_1,$$

which can be rewritten as

$$\iint (\zeta(\boldsymbol{x}_0) - \zeta(\boldsymbol{x}_1))\, \mathrm{H}'_\sigma\left(F(\boldsymbol{x}_1) + c\,\zeta(\boldsymbol{x}_1) - F(\boldsymbol{x}_0) - c\,\zeta(\boldsymbol{x}_0)\right) g_0(\boldsymbol{x}_1) g_1(\boldsymbol{x}_0) \mathrm{d}\boldsymbol{x}_1 \mathrm{d}\boldsymbol{x}_0,$$

by the exchange of $\boldsymbol{x}_0$ for $\boldsymbol{x}_1$ because of the symmetry: $\mathrm{H}'_\sigma(-z) = \mathrm{H}'_\sigma(z)$. Hence, we conclude that

$$\begin{aligned} 2\frac{\partial}{\partial c}\Psi(c) = \iint & (\zeta(\boldsymbol{x}_1) - \zeta(\boldsymbol{x}_0))\, \mathrm{H}'_\sigma\left(F(\boldsymbol{x}_1) + c\zeta(\boldsymbol{x}_1) - F(\boldsymbol{x}_0) - c\,\zeta(\boldsymbol{x}_0)\right) \\ & \times g_0(\boldsymbol{x}_0) g_0(\boldsymbol{x}_1)\, (\Lambda(\boldsymbol{x}_1) - \Lambda(\boldsymbol{x}_0))\, \mathrm{d}\boldsymbol{x}_0 \mathrm{d}\boldsymbol{x}_1, \end{aligned}$$

which is always positive because of the assumption that m is a strictly increasing function. Hence, the function $\Psi(c)$ is strictly increasing.

From the discussion above, it follows that

$$\begin{aligned} \mathrm{AUC}_\sigma(F) &< \lim_{c\to\infty} \Psi(c) \\ &= \lim_{c\to\infty} \mathrm{AUC}_\sigma\left[c\,\left\{\frac{F}{c} + \zeta\right\}\right] \\ &= \lim_{c\to\infty} \mathrm{AUC}_{\frac{\sigma}{c}}\left(\frac{F}{c} + \zeta\right) \\ &= \mathrm{AUC}(\zeta) \\ &= \mathrm{AUC}(\Lambda), \end{aligned}$$

which concludes (5). □

From Theorem 1, we observe that

$$\mathrm{AUC}_\sigma(F) < \mathrm{AUC}(\Lambda),$$

and that no score function $F(\boldsymbol{x})$ can attain the equality above when $\sigma > 0$. Hence, we can perform the supremization of $\mathrm{AUC}_\sigma(F)$ instead of the maximization. This property is not preferable in building an iterative algorithm for maximization of $\mathrm{AUC}_\sigma(F)$; therefore, we propose a regularization scheme for $F(\boldsymbol{x})$ in a subsequent discussion.

3 AUCBoost

3.1 Objective function

We investigate a classification problem based on a boosting method. The key concept is to construct a powerful score function $F(\boldsymbol{x})$ by combining many various weak classifiers (Hastie et al. 2001). Any single weak classifier itself has a very poor ability for classification, whose performance is almost equal to random guessing; however, the combination of a number of them produces a very flexible and strong score function. We aim to construct $F(\boldsymbol{x})$ in such a way based on the AUC.

At first, we prepare a set $\mathcal{F}_k$ for each kth component of $\boldsymbol{x} \in \mathbf{R}^p$:

$$\mathcal{F}_k = \{f(\boldsymbol{x}) = a\mathrm{H}(x_k - b) + (1-a)/2 \mid a \in \{-1, 1\},\ b \in \mathcal{B}_k\}, \quad k = 1, \ldots, p,$$

where $\mathcal{B}_k$ is a finite discrete set, which is determined by taking every intermediate point of samples or a number of sample quantiles. As seen in the definition, $f(\boldsymbol{x})$ is a simple step function taking one of the two values $\{0, 1\}$. Then, we combine the sets into

$$\mathcal{F} = \bigcup_{k=1}^{p} \mathcal{F}_k, \tag{6}$$

called the decision stump class, among which we choose weak classifiers to construct $F(\boldsymbol{x})$. The set $\mathcal{F}$ can be modified to include interaction terms that may improve classification performance. However, the interpretation becomes difficult and unclear especially when the number of feature variables is large. Hence, in this paper we focus only on the main effects of feature variables.

In this setting, $F(\boldsymbol{x})$ can be decomposed as the same way as GAM:

$$\begin{aligned} F(\boldsymbol{x}) &= \sum_{f \in \mathcal{F}'_1} \alpha_f f(\boldsymbol{x}) + \cdots + \sum_{f \in \mathcal{F}'_p} \alpha_f f(\boldsymbol{x}) \\ &= F_1(x_1) + \cdots + F_p(x_p), \end{aligned}$$

where $\mathcal{F}'_k$ is a subset of $\mathcal{F}_k, k = 1, \ldots, p$, whose elements $f's$ are selected in a boosting algorithm in Sect. 3.2, and α_f means a corresponding coefficient of f. Using these notations, the objective function we propose is given as

$$\begin{aligned} \overline{\mathrm{AUC}}_{\sigma,\lambda}(F) = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \mathrm{H}_\sigma(F(\boldsymbol{x}_{1j}) - F(\boldsymbol{x}_{0i})) \\ -\lambda \sum_{k=1}^{p} \sum_{x_k \in \mathcal{B}_k} \left\{F_k^{(2)}(x_k)\right\}^2, \end{aligned} \tag{7}$$

where λ is a smoothing parameter and $F_k^{(2)}(x_k)$ denotes the second-order difference of $F_k(x_k)$: $F_k^{(2)}(x_k) = F_k(x_k^{(-1)}) - 2F_k(x_k) + F_k(x_k^{(+1)})$ with $x_k^{(-1)} < x_k^{(+1)}$. The first term is the approximate empirical AUC based on the standard normal distribution function; the second term gives a penalty for redundant behavior of $F(\boldsymbol{x})$, which focuses on points in $\mathcal{B}_k$ for each k because $F_k(x_k)$ has discontinuities only at the points. Thus, the modeling of $F(\boldsymbol{x})$ is similar to that of GAM. The difference is that the proposed method is based on maximization of the AUC in place of the likelihood, and that we use the second-order difference of $F_k(x_k)$ instead of the second derivative of $F_k(x_k)$ because of its nonsmoothness. The iteration method is also different: we maximize the objective function by a boosting method, whereas GAM is implemented by a backfitting algorithm (Hastie et al. 2001). We investigate the difference in detail using numerical simulation data in Sect. 4.

We note that there is a special relation between the scale parameter σ and the smoothing parameter λ. Equation (7) can be rewritten as

$$\begin{aligned}\overline{\mathrm{AUC}}_{\sigma,\lambda}(F) = \frac{1}{n_0 n_1}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1}\mathrm{H}_\sigma(F(\boldsymbol{x}_{1j}) - F(\boldsymbol{x}_{0i})) \\ -\lambda\sigma^2\sum_{k=1}^{p}\sum_{x_k\in\mathcal{B}_k}\left\{\frac{F_k^{(2)}(x_k)}{\sigma}\right\}^2.\end{aligned}$$

Hence, we have

$$\overline{\mathrm{AUC}}_{\sigma,\lambda}(F) = \overline{\mathrm{AUC}}_{\sigma',\lambda'}\left(\frac{\sigma'}{\sigma}F\right),$$

if $\lambda\sigma^2 = \lambda'\sigma'^2$. This implies that the maximization of $\overline{\mathrm{AUC}}_{\sigma,\lambda}(F)$ is equivalent to that of $\overline{\mathrm{AUC}}_{1,\lambda\sigma^2}\left(\frac{F}{\sigma}\right)$. Therefore, we have

$$\max_{\sigma,\lambda,F}\overline{\mathrm{AUC}}_{\sigma,\lambda}(F) = \max_{\lambda,F}\overline{\mathrm{AUC}}_{1,\lambda}(F).$$

From this consideration, we can fix $\sigma = 1$ without loss of generality. Henceforth, we discuss

$$\overline{\mathrm{AUC}}_{\lambda}(F) = \frac{1}{n_0 n_1}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1}\Phi(F(\boldsymbol{x}_{1j}) - F(\boldsymbol{x}_{0i})) - \lambda\sum_{k=1}^{p}\sum_{x_k\in\mathcal{B}_k}\left\{F_k^{(2)}(x_k)\right\}^2,$$

which is rewrite of $\overline{\mathrm{AUC}}_{1,\lambda}(F)$ for notational convenience. This discussion not only leads to a drastic reduction of the computational cost for the implementation of our method, but also has consistency with Theorem 1. The scale parameter σ, which controls the accuracy of the approximation of the AUC, is not an essential factor in the sense of the supremization of the approximate AUC. On the other hand, the smoothing parameter λ has another important role. As mentioned after Theorem 1, the approximate AUC has no maximum in itself. The penalty term for smoothness in $\mathrm{AUC}_\lambda(F)$

also guarantees the existence of the maximum of $\mathrm{AUC}_\lambda(F)$, and makes the numerical maximization stable.

Ma and Huang (2007) and Wang et al. (2007) approximated the empirical AUC by a sigmoid function, and followed a rule of thumb to determine a scale parameter. That is to say, the accuracy of approximation of the empirical AUC is already fixed before running their algorithm. In contrast, we do not impose such a strict condition; we vary only the smoothing parameter λ and select the best value by cross-validation (see Sect. 3.3).

3.2 AUCBoost algorithm

Let us give a brief explanation of how the score function $F(\boldsymbol{x})$ is constructed by sequentially selecting $f(\boldsymbol{x})$'s in the set $\mathcal{F}$ defined in (6). Our approach is based on a boosting learning algorithm to maximize $\overline{\mathrm{AUC}}_\lambda(F)$ in the linear hull of $\mathcal{F}$, with the number of iterations T.

1. Start with a score function $F_0(\boldsymbol{x})$.
2. For $t = 1, \ldots, T$
   a. Find the best weak classifier f_t and calculate the coefficient α_t as

$$f_t(\boldsymbol{x}) = \operatorname*{argmax}_{f \in \mathcal{F}} \frac{\partial}{\partial \alpha} \overline{\mathrm{AUC}}_\lambda(F_{t-1} + \alpha f)\Big|_{\alpha=0},$$
$$\alpha_t = \operatorname*{argmax}_{\alpha > 0} \overline{\mathrm{AUC}}_\lambda(F_{t-1} + \alpha f_t).$$

   b. Update the score function as

$$F_t(\boldsymbol{x}) = F_{t-1}(\boldsymbol{x}) + \alpha_t f_t(\boldsymbol{x}).$$

3. Finally, output the final score function:

$$F(\boldsymbol{x}) = F_0(\boldsymbol{x}) + \sum_{t=1}^{T} \alpha_t f_t(\boldsymbol{x}).$$

If we have no prior information about the data, we set $F_0(\boldsymbol{x}) = 0$. In step 2.a, we search $\mathcal{F}$ for a $f_t(\boldsymbol{x})$ which maximizes the first derivative of $\overline{\mathrm{AUC}}_\lambda(F)$ at the point $F_{t-1}(\boldsymbol{x}) + \alpha f(\boldsymbol{x})$. This argument is similar to that of Hastie et al. (2001) and Takenouchi and Eguchi (2004). Next, we calculate the coefficient of $f_t(\boldsymbol{x})$ using the Newton–Raphson method, and add $\alpha_t f_t(\boldsymbol{x})$ to the previous score function. We repeat this process T times and output the final score function. Thus, the resultant score function is an aggregation of $f_t(\boldsymbol{x})$'s with weights α_t's. Further details of this algorithm are as follows.

In step 2.a, we search $\mathcal{F}$ for f_t that satisfies

$$
\begin{aligned}
f_t(\boldsymbol{x}) &= \underset{f\in\mathcal{F}}{\operatorname{argmax}} \frac{\partial}{\partial\alpha}\overline{\mathrm{AUC}}_\lambda(F_{t-1}+\alpha f)\Big|_{\alpha=0} \\
&= \underset{\substack{k\in\{1,\ldots,p\}\\ a\in\{-1,1\}\\ b\in\mathcal{B}_k}}{\operatorname{argmax}} \frac{1}{n_0n_1}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1}\phi\left(F_{t-1}(\boldsymbol{x}_{1j})-F_{t-1}(\boldsymbol{x}_{0i})\right) \\
&\quad\times\{a\mathrm{H}(x_{1jk}-b)-a\mathrm{H}(x_{0ik}-b)\} \\
&\quad-2\lambda\sum_{x_k\in\mathcal{B}_k}\left\{F_k(x_k^{(-1)})-2F_k(x_k)+F_k(x_k^{(+1)})\right\} \\
&\quad\times\left\{a\mathrm{H}(x_k^{(-1)}-b)-2a\mathrm{H}(x_k-b)+a\mathrm{H}(x_k^{(+1)}-b)\right\},
\end{aligned}
$$

where ϕ is the standard normal density function, $F_k(x_k)$ is the kth component of $F_{t-1}(\boldsymbol{x})$ (a score function of x_k at an iteration number $t-1$), and x_{0ik}, x_{1jk} are the kth component of $\boldsymbol{x}_{0i}, \boldsymbol{x}_{1j}$, respectively.

Then, the second term in the equation above is calculated into

$$
\begin{aligned}
&-2\lambda\left[\left\{F_k(b^{(-2)})-2F_k(b^{(-1)})+F_k(b)\right\}a-\left\{F_k(b^{(-1)})-2F_k(b)+F_k(b^{(+1)})\right\}a\right] \\
&\quad=-2\lambda a\left\{F_k(b^{(-2)})-3F_k(b^{(-1)})+3F_k(b)-F_k(b^{(+1)})\right\},
\end{aligned}
$$

where an element with a smaller superscript number than that of the minimum element in $\mathcal{B}_k$ is set to the minimum one. Similarly, an element with a larger superscript number than that of the maximum element is set to the maximum one. In regard to the coefficient (α_t) of $f_t(\boldsymbol{x})$, we seek it by the Newton–Raphson method using

$$
\begin{aligned}
&\frac{\partial}{\partial\alpha}\overline{\mathrm{AUC}}_\lambda(F_{t-1}+\alpha f_t) \\
&\quad=\frac{1}{n_0n_1}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1}\phi\big(F_{t-1}(\boldsymbol{x}_{1j})-F_{t-1}(\boldsymbol{x}_{0i})+\alpha\{f_t(\boldsymbol{x}_{1j}) \\
&\qquad-f_t(\boldsymbol{x}_{0i})\}\big)\left(f_t(\boldsymbol{x}_{1j})-f_t(\boldsymbol{x}_{0i})\right) \\
&\qquad-2\lambda\left[a_t\left\{F_{k_t}(b_t^{(-2)})-3F_{k_t}(b_t^{(-1)})+3F_{k_t}(b_t)-F_{k_t}(b_t^{(+1)})\right\}+2\alpha\right],
\end{aligned}
$$

and

$$
\begin{aligned}
&\frac{\partial^2}{\partial\alpha^2}\overline{\mathrm{AUC}}_\lambda(F_{t-1}+\alpha f_t) \\
&\quad=-\frac{1}{n_0n_1}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1}\phi\big(F_{t-1}(\boldsymbol{x}_{1j})-F_{t-1}(\boldsymbol{x}_{0i})+\alpha\{f_t(\boldsymbol{x}_{1j}) \\
&\qquad-f_t(\boldsymbol{x}_{0i})\}\big)\left(f_t(\boldsymbol{x}_{1j})-f_t(\boldsymbol{x}_{0i})\right)^2 \\
&\qquad\times\left(F_{t-1}(\boldsymbol{x}_{1j})-F_{t-1}(\boldsymbol{x}_{0i})+\alpha\{f_t(\boldsymbol{x}_{1j})-f_t(\boldsymbol{x}_{0i})\}\right)-4\lambda, \qquad (8)
\end{aligned}
$$

where

$$f_t(\boldsymbol{x}) = a_t \, \mathrm{H}(x_{k_t} - b_t) + (1 - a_t)/2.$$

The first term in (8) is usually negative for an appropriate value of α. However, it happens to be positive in the Newton–Raphson process. Our objective is to obtain α that maximizes $\overline{\mathrm{AUC}}_\lambda(F_{t-1} + \alpha f_t)$, so the sign of (8) should be always negative. We find that the smoothing parameter λ stabilizes the algorithm of AUCBoost.

3.3 Tuning parameter selection

In our method there are two parameters to be determined: a smoothing parameter λ and the iteration number T. We use the following K-fold cross-validation. At first, we partition the whole data set into K subsets of almost equal sizes, and evaluate an objective function such as

$$\mathrm{AUC}_{\mathrm{CV}}(\lambda, T) = \frac{1}{K} \sum_{i=1}^{K} \overline{\mathrm{AUC}}_\lambda^{(i)}(F^{(-i)}),$$

where $F^{(-i)}$ is a score function constructed by AUCBoost using the data set without the ith subset, and $\overline{\mathrm{AUC}}_\lambda^{(i)}$ is the $\overline{\mathrm{AUC}}_\lambda$ calculated on the ith subset alone. We give a typical example of results of $\mathrm{AUC}_{\mathrm{CV}}$ in Fig. 2, assuming $\boldsymbol{X}_0 \sim \mathrm{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and $\boldsymbol{X}_1 \sim \mathrm{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, where $\boldsymbol{\mu}_0 = (0, 0, 0, 0)'$, $\boldsymbol{\mu}_1 = (0, 0.5, 0, 0.5)'$, $\boldsymbol{\Sigma}_0 = \mathrm{diag}(1, 1, 1, 1)$ and $\boldsymbol{\Sigma}_1 = \mathrm{diag}(1, 1, 4, 0.25)$. In this case, the best pair of the parameters seems to be $\lambda = 0.01$ and $T = 200$. The curve with $\lambda = 0.0001$ increases rapidly at the beginning and starts to decline around $T = 20$. The second curve denoted by triangles has a peak around $T = 70$ and shows a moderate tendency to decrease after that point. On the other hand, the best curve with $\lambda = 0.01$ shows that the score function hardly suffers from overfitting to the data. This fact also can be confirmed by observing the corresponding score function $F(\boldsymbol{x})$. The true score function in this setting is a smooth function; however, we observed that the score function with $\lambda = 0.0001$ clearly lacked the smoothness (not shown here). This also indicates overfitting to the data. With an appropriate value of λ and a relatively large iteration number T, this slow learning process contrasts starkly with the usual regularization technique, i.e., early stopping (Zhang and Yu 2005). We set the value of K to 10 for simulation studies and 5 for a real data analysis, according to the sample size.

3.4 Score plot and score ROC

We discuss the AUCBoost algorithm to select classifiers in the decision stump class $\mathcal{F}$. The choice of the class provides us with useful information regarding the feature

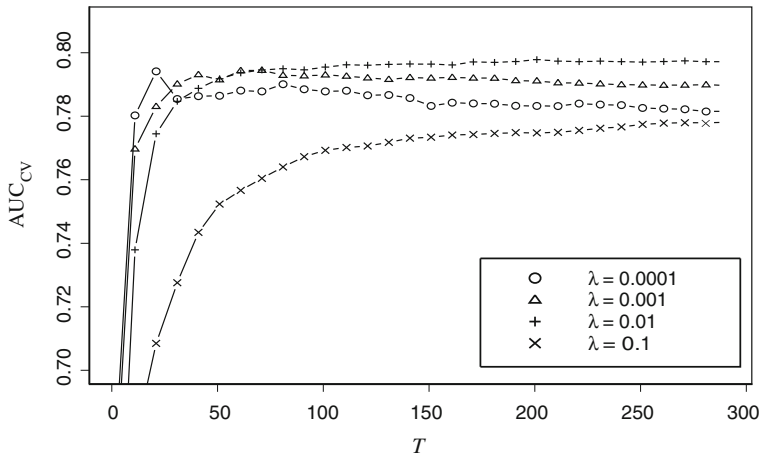

Fig. 2 Results of $\mathrm{AUC}_{\mathrm{CV}}$ corresponding to different values of λ, as a function of the number of iterations T

variables in a post-analysis of classification. The final score function $F(\boldsymbol{x})$ is decomposed as

$$F(\boldsymbol{x}) = \sum_{k=1}^{p} F_k(x_k).$$

The utility of the plot of $F_k(x_k)$ against x_k (score plot of x_k) is referred to by Friedman et al. (2000) and Kawakita et al. (2005). Observing each score plot very carefully, we are able to not only understand how each feature variable x_k influences the classification performance, but also know which feature variable is the most effective and informative one. We discuss this utility more in detail in simulation studies. Another useful way to gauge the efficiency of each feature variable is to draw the ROC curve for $F_k(x_k)$ (score ROC) and calculate the corresponding AUC (score AUC). $F_k(x_k)$ represents the contribution of x_k to the total classification performance; hence, the value of the score AUC shows the utility of x_k. These measurements are more convenient for comparing the utilities of feature variables because we can order them according to their values.

4 Simulation studies

4.1 Setting

In this section, we present two simulation studies. One is intended to demonstrate that the score function $F(\boldsymbol{x})$ generated by AUCBoost provides a good approximation to the optimal score function, and that score plots are useful for evaluating each feature variable's contribution to $F(\boldsymbol{x})$. The other is designed to show that, in cases where several outliers exist, AUCBoost is much more powerful and robust than other classification methods such as AdaBoost, GAM and the generalized linear model (GLM).

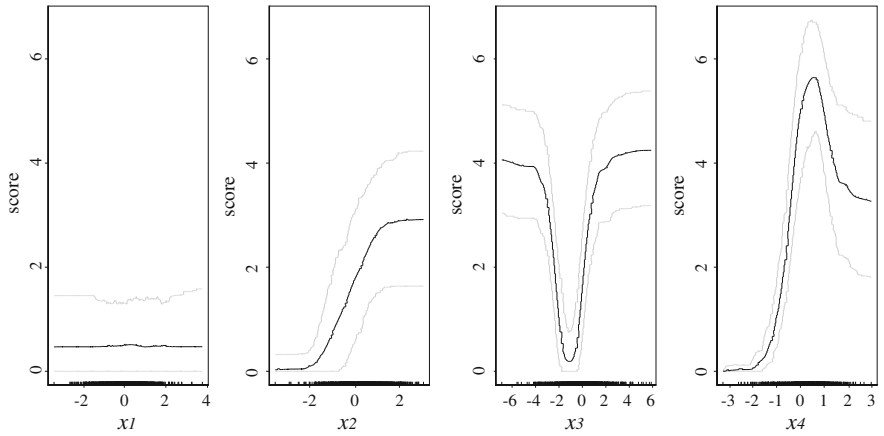

Fig. 3 Score plots for AUCBoost. The *black lines* indicate mean score plots and the *gray lines* indicate the 95% pointwise confidence bands

The iteration number for AdaBoost is also determined by cross-validation where the objective function is based on the empirical AUC. Cubic splines are used for GAM, and these simulation studies are done using Splus 8.0. Throughout these simulations, the training sample size is set to be 500 ($n_0 = 250$, $n_1 = 250$) and we evaluated the quality using a test sample of size 200 ($n_0 = 100$ $n_1 = 100$). Summary statistics are based on 1,000 repetitions.

4.2 Comparison with the optimal score function

Consider the same situation as that of Sect. 3.3: $\boldsymbol{X}_0 \sim \mathrm{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and $\boldsymbol{X}_1 \sim \mathrm{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, where $\boldsymbol{\mu}_0 = (0, 0, 0, 0)'$, $\boldsymbol{\mu}_1 = (0, 0.5, 0, 0.5)'$, $\boldsymbol{\Sigma}_0 = \mathrm{diag}(1, 1, 1, 1)$ and $\boldsymbol{\Sigma}_1 = \mathrm{diag}(1, 1, 4, 0.25)$. From Eq. (4), the optimal score function in this setting is given as

$$F_{\mathrm{N}}(\boldsymbol{x}) = \boldsymbol{x}'(\boldsymbol{\Sigma}_0^{-1} - \boldsymbol{\Sigma}_1^{-1})\boldsymbol{x} + 2(\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_0'\boldsymbol{\Sigma}_0^{-1})\boldsymbol{x},$$

which coincides with a linear score function proposed by Su and Liu (1993) if $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1$. The score plots constructed by AUCBoost track $F_{\mathrm{N}}(\boldsymbol{x})$ very well as seen in Fig. 3, where the rug plots at the bottom of each graph depict the data distribution. Clearly, it shows nonlinearity of $F(\boldsymbol{x})$, especially $F_3(x_3)$ and $F_4(x_4)$. From the shape of $F_3(x_3)$ we see that x_3 with class label 0 has a tendency to concentrate around the origin, compared to x_3 with class label 1. On the other hand, in regard to $F_4(x_4)$, we see the opposite tendency of x_4. The flatness of $F_1(x_1)$ means that x_1 is useless for discriminating subjects with class 0 from those with class 1, because weak classifiers for x_1 are rarely chosen, and the weight coefficients are calculated to be very small in the AUCBoost algorithm. Judging from the heights of score plots, x_4 seems to be the most informative one.

Table 1 shows the results of the score AUCs and the AUCs calculated by AUCBoost and $\hat{F}_{\mathrm{N}}$, where $\hat{F}_{\mathrm{N}}$ denotes the estimator of F_{N}. As expected, $\hat{F}_{\mathrm{N}}$ achieves superior

Table 1 The mean score AUCs and the AUCs with 95% confidence bands in parentheses

	x_1	x_2	x_3	x_4	Total
AUCBoost	0.501	0.628	0.700	0.736	0.828
	(0.429, 0.579)	(0.548, 0.707)	(0.617, 0.774)	(0.670, 0.799)	(0.772, 0.879)
$\hat{F}_{\mathrm{N}}$	0.500	0.638	0.703	0.742	0.840
	(0.425, 0.583)	(0.565, 0.714)	(0.633, 0.779)	(0.668, 0.809)	(0.780, 0.887)

performance for all AUCs. It is because $\hat{F}_{\mathrm{N}}$ is derived based on the underlying probability distributions; on the other hand, the score function of AUCBoost is constructed by the sample distributions. We also notice that the values of the score AUCs for AUCBoost are in accordance with the heights of the score plots. The utility of x_4 is confirmed again.

4.3 Comparison with other methods

Next, we relax the conditions of the probability distribution a little, and consider a multivariate t-distribution. This is a more practical setting because it contains several outliers which we often observe in real data. While there are several forms of multivariate t-distribution, we use the most common one. The density function of p-dimensional t-distribution with ν degrees of freedom, mean vector $\boldsymbol{\mu}$ and precision matrix $\boldsymbol{\Sigma}^{-1}$, is given as

$$g(\boldsymbol{x}) = \frac{\Gamma(\frac{p+\nu}{2})\sqrt{|\boldsymbol{\Sigma}^{-1}|}}{\Gamma(\frac{\nu}{2})(\nu\pi)^{\frac{p}{2}}}\left[1+\frac{1}{\nu}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]^{-\frac{p+\nu}{2}}.$$

We use the same parameters as those in the previous subsection: $\boldsymbol{\mu}_0 = (0,0,0,0)'$, $\boldsymbol{\mu}_1 = (0,0.5,0,0.5)'$, $\boldsymbol{\Sigma}_0 = \mathrm{diag}(1,1,1,1)$ and $\boldsymbol{\Sigma}_1 = \mathrm{diag}(1,1,4,0.25)$. To focus on the investigation of the robustness of $F(\boldsymbol{x})$ constructed by AUCBoost, we consider an extreme situation ($\nu = 1$). Figure 4 shows score plots and score ROCs of x_3 for AUCBoost, AdaBoost, GAM and GLM. The range of score plots has been adjusted for a better view. Interestingly, the shape of score plots of AUCBoost and AdaBoost are almost the same. This is because both of the boosting methods focus only on points that are useful for the classification. On the other hand, GAM is sensitive to uninformative samples such as outliers, which causes the GAM's performance instability (Kawakita et al. 2005). In regard to GLM, it does not capture the useful information about x_3 at all, which is observable from the value of the score AUC (0.496) as well as the shape of the score ROC. The concavity of the shape of ROC is known to be a necessary condition of the optimality (Pepe 2003). In the last column of Table 2, we can see that the corresponding 95% confidence band of the AUC for AUCBoost is much narrower than the others. Among all of them, the result for AUCBoost is the most stable with the largest mean AUC value (0.787). The smoothing parameter λ contributes to the stable result of AUCBoost.

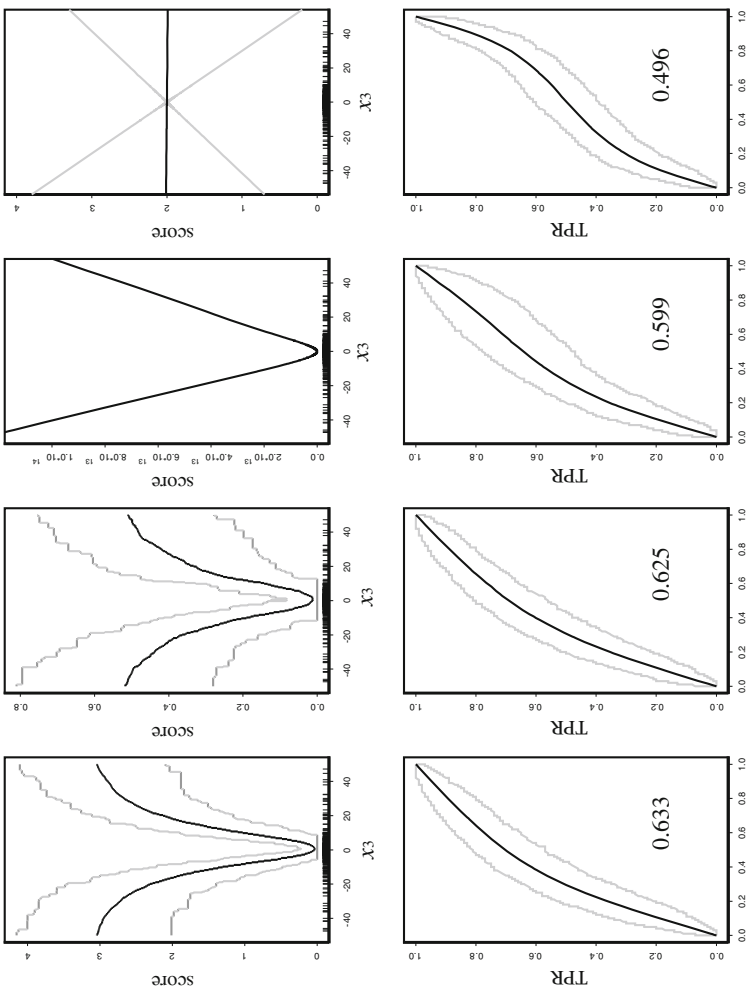

Fig. 4 Results of score plots (*upper panels*) and score ROCs (*lower panels*) for x_3 of AUCBoost, AdaBoost, GAM and GLM. The *black lines* indicate mean score plots and score ROCs, and the *gray lines* indicate the 95% pointwise confidence bands. The confidence band of the score plot for GAM is omitted, and the minimum values of axes of score plots are set to 0 for a better view

Table 2 The mean score AUCs and the AUCs with 95% confidence bands in parentheses

	x_1	x_2	x_3	x_4	Total
AUCBoost	0.501	0.616	0.633	0.697	0.787
	(0.425, 0.578)	(0.541, 0.694)	(0.547, 0.711)	(0.620, 0.764)	(0.723, 0.839)
AdaBoost	0.501	0.601	0.625	0.690	0.776
	(0.428, 0.579)	(0.522, 0.685)	(0.553, 0.696)	(0.610, 0.761)	(0.705, 0.834)
GAM	0.497	0.618	0.599	0.672	0.738
	(0.418, 0.583)	(0.543, 0.699)	(0.487, 0.694)	(0.603, 0.748)	(0.661, 0.804)
GLM	0.501	0.601	0.496	0.649	0.648
	(0.434, 0.572)	(0.388, 0.690)	(0.427, 0.555)	(0.340, 0.744)	(0.533, 0.737)

5 Application to spinal disease in children data

We apply AUCBoost to a real data set, which can be seen in Statistical Models in S edited by Chambers and Hastie (1992). The label is the outcome of corrective spinal surgery of 81 children: whether kyphosis is present or absent. The feature variables are as follows: Age, the age of the child in months; Number, the number of vertebrae in the operation; and Start, the beginning of the range of vertebrae involved in the operation. We used the first 70 samples as training data, and the others as test data. Figure 5 shows the score plots for AUCBoost, AdaBoost, GAM and GLM, respectively. We find clear nonlinearity of score plots for Age, except for that of GLM. The peak appears around 100 months. A child of Age 200 is estimated by GLM to have the highest risk of a postoperative deformity; on the other hand, the risk at this age is estimated by GAM to be the lowest. AUCBoost gives results intermediate between these two extremes. The smoothness of score plots for AUCBoost is quite different from that of AdaBoost. This result makes it easy to understand how each feature variable affects the outcome after surgery and to interpret the results of the analysis. It also contributes to preventing the score function $F(\boldsymbol{x})$ from overfitting to the data. The values of the AUCs based on training data are 0.926, 0.997, 0.949 and 0.869 for AUCBoost, AdaBoost, GAM and GLM; however, the values based on test data are 0.777, 0.666, 0.666 and 0.666, respectively.

6 Conclusions and future work

AUCBoost offers a flexible combination of multiple feature variables, which is optimal in the sense of the maximization of the AUC. The smoothing parameter λ in the objective function not only contributes to improvement of the classification performance, but also gives us smoothed score plots, which are very useful in clinical studies. By observing the score plots very carefully, we can understand how each feature variable is associated with a disease or other endpoint, and also evaluate its efficiency by calculating the corresponding AUC.

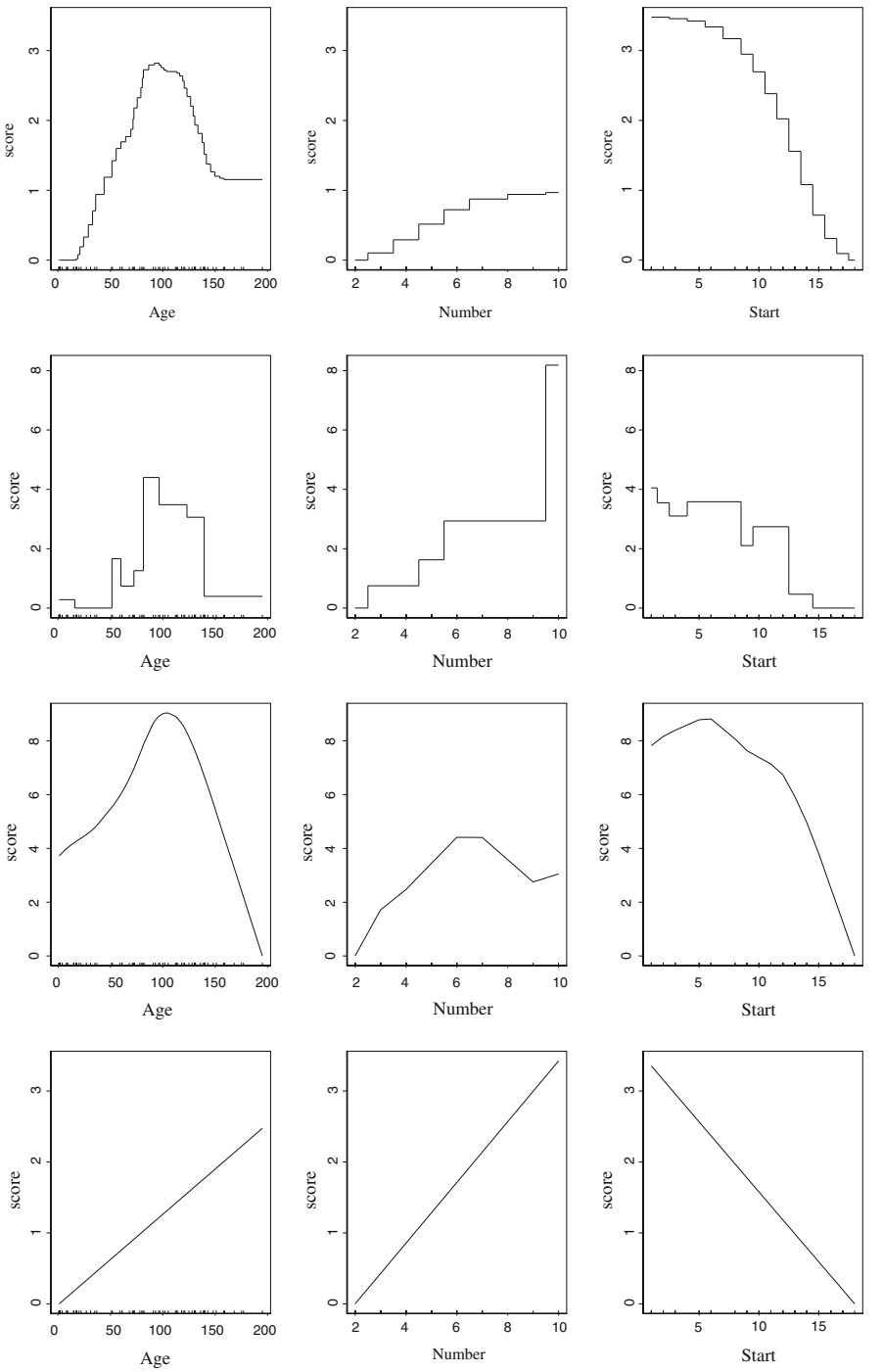

Fig. 5 Score plots for AUCBoost, AdaBoost, GAM and GLM from top to bottom. The minimum values of each score plot are set to 0 for better view

From the setting of $\mathcal{F}$ which consists of component-based simple classifiers, the score function of AUCBoost has a similar form to that of GAM. However, there are two major differences between them. First, we maximize the AUC instead of the likelihood. Second, we update the score function by sequentially adding weak classifiers, whereas GAM is based on a backfitting algorithm (Hastie et al. 2001). The forward stagewise additive modeling gives AUCBoost robustness to distributions of data as seen in Sect. 4.3. Thus, AUCBoost is expected to show stable classification performance in various situations. This property also makes it easy to take discrete or ordered categorical data into consideration, which is difficult or impossible for the backfitting algorithm.

A weak point of AUCBoost is that the selection of the tuning parameter λ and T is time-consuming because we apply a simple cross-validation method. In order to avoid such a computational cost and make it easy to use, a more sophisticated procedure is necessary. Recently, Ueki and Fueda (2009) proposed an effective method for determining tuning parameters of maximum penalized likelihood estimator. The idea is based on likelihood, not the AUC; however, it could be modified into AUCBoost and help it reduce its computational costs.

AUCBoost can also be applied to a high-dimensional data analysis, in which variable selection is much more important than in the low-dimensional data analysis we consider in this paper. The AUCBoost algorithm implicitly includes a selection process at each iteration stage, so that informative feature variables are selected as a result after applying AUCBoost. This property is similar to GAMBoost (Tutz and Binder 2006), which circumvents GAM's restriction to low-dimensional setting. The concept of the partial AUC (pAUC) is also of great interest in the analysis of genetic data. Pepe et al. (2003) showed the biological utility of the pAUC for ranking informative genes. We will work on developing partial AUCBoost as one of the appealing extensions of AUCBoost.

Acknowledgments This study was supported by the Program for Promotion of Fundamental Studies in Health Sciences of the National Institute of Biomedical Innovation (NIBIO).

References

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology, 12*, 387–415.

Chambers, J. M., Hastie, T. J. (1992). *Statistical models in S*. Pacific Grove, CA: Wadsworth and Brooks.

Eguchi, S., Copas, J. (2002). A class of logistic-type discriminant functions. *Biometrika, 89*, 1–22.

Freund, Y., Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences, 55*, 119–139.

Friedman, J., Hastie, T., Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting (with discussion). *The Annals of Statistics, 28*, 337–407.

Hastie, T., Tibshirani, R. (1986). Generalized additive models. *Statistical Science, 1*, 297–318.

Hastie, T., Tibshirani, R., Friedman, J. (2001). *The elements of statistical learning*. New York: Springer.

Kawakita, M., Minami, M., Eguchi, S., Lennert-Cody, C. E. (2005). An introduction to the predictive technique AdaBoost with a comparison to generalized additive models. *Fisheries Research, 76*, 328–343.

Long, P. M., Servedio, R. A. (2007). Boosting the area under the ROC curve. In J. C. Platt, D. Koller, Y. Singer, S. Roweis (Eds.), *Advances in neural information processing systems* (Vol. 20, pp. 945–952). Cambridge, MA: MIT Press.

Ma, S., Huang, J. (2005). Regularized ROC method for disease classification and biomarker selection with microarray data. *Bioinformatics, 21*, 4356–4362.

Ma, S., Huang, J. (2007). Combining multiple markers for classification using ROC. *Biometrics, 63*, 751–757.

McIntosh, M. W., Pepe, M. S. (2002). Combining several screening tests: Optimality of the risk score. *Biometrics, 58*, 657–664.

McLachlan, G. J. (2004). *Discriminant analysis and statistical pattern recognition.* New York: Wiley.

Murata, N., Takenouchi, T., Kanamori, T., Eguchi, S. (2004). Information geometry of $\mathcal{U}$-Boost and Bregman divergence. *Neural Computation, 16*, 1437–1481.

Neyman, J., Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A, 231*, 289–337.

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction.* Oxford: Oxford University Press.

Pepe, M. S., Thompson, M. L. (2000). Combining diagnostic test results to increase accuracy. *Biostatistics, 1*, 123–140.

Pepe, M. S., Longton, G., Anderson, G. L., Schummer, M. (2003). Selecting differentially expressed genes from microarray experiments. *Biometrics, 59*, 133–142.

Pepe, M. S., Cai, T., Longton, G. (2006). Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics, 62*, 221–229.

Pepe, M. S., Janes, H., Longton, G., Leisenring, W., Newcomb, P. (2004). Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American Journal of Epidemiology, 159*, 882–890.

Su, J. Q., Liu, J. S. (1993). Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association, 88*, 1350–1355.

Takenouchi, T., Eguchi, S. (2004). Robustifying AdaBoost by adding the naive error rate. *Neural Computation, 16*, 767–787.

Tutz, G., Binder, H. (2006). Generalized additive modeling with implicit variable selection by likelihood-based boosting. *Biometrics, 62*, 961–971.

Ueki, M., Fueda, K. (2009). Optimal tuning parameter estimation in maximum penalized likelihood method. *Annals of the Institute of Statistical Mathematics*. doi:10.1007/s10463-008-0186-0.

Wang, Z., Chang, Y. I., Ying, Z., Zhu, L., Yang, Y. (2007). A parsimonious threshold-independent protein feature selection method through the are under receiver operating characteristic curve. *Bioinformatics, 23*, 2788–2794.

Zhang, B. T., Yu, B. (2005). Boosting with early stopping: Convergence and consistency. *The Annals of Statistics, 33*, 1538–1579.

Zhou, X. H., Obuchowski, N. A., McClish, D. K. (2002). *Statistical methods in diagnostic medicine.* New York: Wiley.