Communications of the Korean Statistical Society
2009, Vol. 16, No. 6, 1005–1012

The Unified Framework for AUC Maximizer

Jongjun Jun[a], Yongdai Kim[a], Sang-Tae Han[b], Hyuncheol Kang[b], Hosik Choi[1,b]

[a]Department of Statistics, Seoul National University
[b]Department of Informational Statistics, Hoseo University

Abstract

The area under the curve(AUC) is commonly used as a measure of the receiver operating characteristic(ROC) curve which displays the performance of a set of binary classifiers for all feasible ratios of the costs associated with true positive rate(TPR) and false positive rate(FPR). In the bipartite ranking problem where one has to compare two different observations and decide which one is "better", the AUC measures the quantity that ranking score of a randomly chosen sample in one class is larger than that of a randomly chosen sample in the other class and hence, the function which maximizes an AUC of bipartite ranking problem is different to the function which maximizes (minimizes) accuracy (misclassification error rate) of binary classification problem. In this paper, we develop a way to construct the unified framework for AUC maximizer including support vector machines based on maximizing large margin and logistic regression based on estimating posterior probability. Moreover, we develop an efficient algorithm for the proposed unified framework. Numerical results show that the proposed unified framework can treat various methodologies successfully.

Keywords: ROC curve, AUC, bipartite ranking problem.

1. Introduction

The area under the curve(AUC) is commonly used as a measure of the receiver operating characteristic(ROC) curve which displays the performance of a set of binary classifiers for all feasible ratios of the costs associated with true positive rate(TPR) and false positive rate(FPR). For evaluating the performance of a classifier, it has become a good alternative to accuracy. The AUC is a quantity which is the ratio of cases that a ranking score of a randomly chosen sample in one class is larger than that of a randomly chosen sample in the other class. It is equivalent to Mann-Whitney statistics. If all samples of one class are ranked higher than all sample of the other class, the AUC achieves 1, which means a perfect ranking. In this view, naive AUC based on the ROC curve which is constructed from varying the threshold (intercept) via usual linear classification function is different to AUC which is directly found from bipartite ranking function in that different classification functions with the same error rate may produce ranking functions with very different AUC values.

The main objective of the bipartite ranking problem is to find relative order not absolute relevance score. Such differences between the binary classification problem and the bipartite ranking problem are theoretically justified by researches including Cortes and Mohri (2004), Agarwal *et al.* (2005) and Clémencon *et al.* (2006) to name just a few. Agarwal *et al.* (2005) derived the distribution-free probabilistic bounds on the deviation of the empirical AUC of a ranking function. Clémencon *et al.* (2006) proved the Fisher consistency of ranking function with respect to 0–1 loss function and furthermore, they proved that ranking function using surrogate loss function of 0–1 loss function is

[1] Corresponding author: Full Time Lecturer, Department of Informational Statistics and Institute of Basic Science, Hoseo University, Asan Campus, Sechul-ri, Baebang-myun, Asan, Chungnam 336-795, Korea. E-mail: choi.hosik@gmail.com.

Bayes consistent. Cortes and Mohri (2004) proved that more specific connections between ranking function and classification function according to ratio of two groups.

Many learning algorithms are proposed by various researchers. Freund *et al.* (2003) proposed the Rankboost using the exponential loss function as the surrogate of 0–1 loss function of two rankers. Joachims (2002) considered that how to learn ranking from click-through data using ranking support vector machines(SVM) model. Brefeld and Scheffer (2005) showed that their ranking formulation is equivalent to one-class support vector machine and furthermore, they reduced computational burden using small clustered sample. Bach *et al.* (2006) showed that the ROC curve obtained by varying both the intercept and the cost asymmetry (slope) always outperforms usual ROC curve obtained by varying only the intercept.

In this paper, we consider an unified framework in the AUC maximization problem. Under the proposed unified framework, we can easily compare various methodologies including SVM based on maximizing large margin and logistic regression based on estimating posterior probability. Moreover, the hybrid methods of such methods can be acquired in the sense that hybrid methods can maximize large margin and estimate posterior simultaneously.

The paper is organized as follows. In Section 2, we review the AUC maximization including ROC curve and then propose the unified framework for AUC maximization. Also, an optimization algorithm is presented. Numerical results on simulated data set are presented in Section 3. Concluding remarks follow in Section 4.

2. Methodology

2.1. ROC and AUC: Review

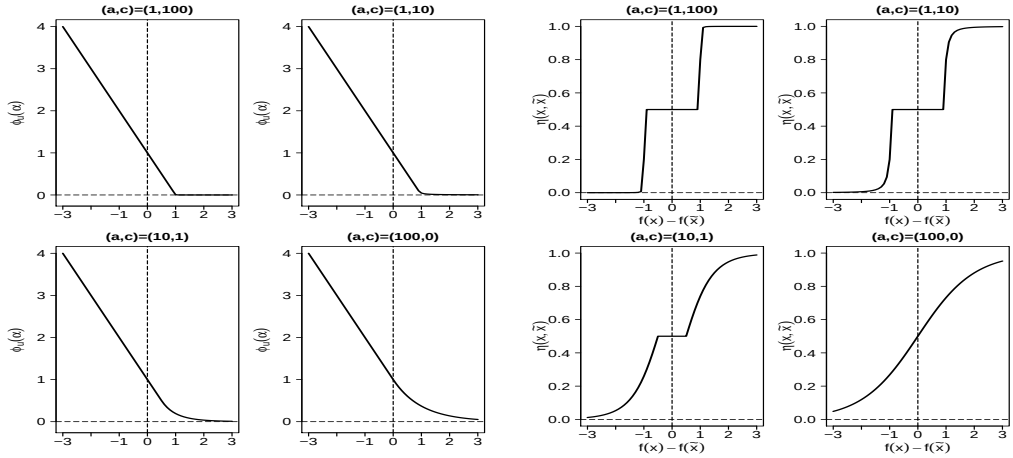
Let $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ be input-output pairs of given data where $\mathbf{x}_i \in \mathbb{R}^p (\triangleq \mathcal{X})$ is a covariate random vector and $y_i \in \{+1, -1\}$ denotes a class label. We assume that there are n_+ positive samples and n_- negative samples. Denote a random vector by (X, Y). And let $(\tilde{X}, \tilde{Y})$ be an independent copy of (X, Y). Denote I_+ and I_- by indices containing the "+1" class points and the "−1" class points respectively.

The goal of bipartite ranking problem is to learn that the rank of X of $Y = 1$ is higher than that of $\tilde{X}$ of $\tilde{Y} = -1$ as much as possible. To clarify this, we introduce a ranking rule which is defined by $r : \mathcal{X} \times \mathcal{X} \to \{-1, 1\}$. The performance of a ranking rule is measured by the ranking risk $L(r) = \Pr(Z \cdot r(X, \tilde{X}) < 0)$ where $Z = (Y - \tilde{Y})/2$, where $(X, Y) = (X, 1)$ and $(\tilde{X}, \tilde{Y}) = (\tilde{X}, -1)$.

Let $f : \chi \longrightarrow \mathbb{R}$ be a scoring function. Given f, $\mathrm{ROC}(f)$ is defined by plotting the $\mathrm{TPR}_f(u) = \Pr(f(X) > u | Y = 1)$ against the $\mathrm{FPR}_f(u) = \Pr(f(X) > u | Y = -1)$ varying $u \in \mathbb{R}$. Let the $(1 - \alpha)^{th}$ quantile of $f(X)$ given $Y = -1$ be $q_{f,\alpha} = \inf_{u \in \mathbb{R}} \{u : \mathrm{FPR}_f(u) \leq \alpha\}$. Then the power at level α is $\mathrm{TPR}_f(q_{f,\alpha}) = \Pr(f(X) > q_{f,\alpha} | Y = 1)$. Clémencon *et al.* (2006) and Clémencon *et al.* (2008) verified that $\mathrm{AUC}(f)$ may be interpreted in a probabilistic fashion via following relation:

$$\begin{aligned}
\mathrm{AUC}(f) &= \int_0^1 \mathrm{TPR}_f(q_{f,\alpha}) d\alpha = \int_0^1 \Pr\left(f(X) > q_{f,\alpha} | Y = 1\right) d\alpha \\
&= E\left(\Pr\left(f(X) > F_f^{-1}(U) | Y = 1\right)\right) \quad \text{where } U \sim \mathrm{Unif}(0,1) \text{ indep. of } (X, Y) \\
&= \Pr\left(f(X) > f(\tilde{X}) | Y = 1, \tilde{Y} = -1\right) \\
&= 1 - \frac{1}{\Pr(Y = 1) \Pr\left(\tilde{Y} = -1\right)} \Pr\left(Z \cdot r(X, \tilde{X}) < 0\right), \qquad (2.1)
\end{aligned}$$

where F_f is the distribution function of $f(X)$ given $Y = -1$ and F_f^{-1} is the inverse function of F_f.



(a) Unified loss functions (b) Corresponding probabilities

Figure 1: *The unified loss functions and corresponding probabilities for various c and a values.*

From Equation (2.1), maximizing AUC(f) is equivalent to minimizing misranking error rate

$$\Pr\left(Z \cdot r\left(X, \tilde{X}\right) < 0\right) = EI\left[\left(Y - \tilde{Y}\right)\left(f(X) - f(\tilde{X})\right) < 0\right], \tag{2.2}$$

where $I[\cdot]$ is the indicator function, which assumes 1 if its argument is true, and 0 otherwise.

A traditional algorithm of learning ranking function tries to find the optimal hyperplane minimizing the empirical risk of Equation (2.2), $1/(n_+n_-)\sum_{i\in I_+, j\in I_-} I(f(\mathbf{x}_i) < f(\mathbf{x}_j))$ among all linear hyperplanes $f(\mathbf{x}) = \beta_0 + \mathbf{x}'\beta, \beta_0 \in \mathbb{R}$ and $\beta \in \mathbb{R}^p$. However, it is known that this estimation is unstable. Powerful alternatives are to shrink solution as well as enhance the ability of model's explanation simultaneously (*e.g*, Tibshirani, 1996). The general form about methods of estimating the optimal hyperplane is to minimize a regularized empirical risk given as

$$\frac{1}{n_+n_-}\sum_{i\in I_+, j\in I_-} \phi\left(-\left(f(\mathbf{x}_i) - f(\mathbf{x}_j)\right)\right) + J_\lambda(\beta), \tag{2.3}$$

where $\phi(z)$ is a convex surrogate loss function of the 0–1 loss $I(\cdot)$ in Equation (2.2) and J_λ is a penalty function where λ controls misranking error rate and function's complexity.

The various methodologies which are inherited from classification methods are proposed: RankBoost using exponential loss function $\phi(z) = \exp(-z)$ or $\phi(z) = \log(1 + \exp(-z))$ by Freund *et al.* (2003), ranking SVM using hinge loss function $\phi(z) = (1 - z)_+$ where $(z)_+ = \max\{z, 0\}$ by Joachims (2002).

2.2. The unified framework in bipartite ranking problem

In binary classification setting, Liu and Zhang (2009) presented the unified large margin machine which can produce logistic regression and SVM (Cortes and Vapnik, 1995), and their hybrid versions. The main advantage of the method is that various methodologies can be easily compared under the unified framework. In this section, following the spirit of Liu and Zhang (2009), we are to propose the unified framework for AUC maximization problem.

Now, the unified loss function is defined as follows:

$$\phi_{\mathrm{u}}(z) = \begin{cases} 1 - z, & \text{if } z \le \dfrac{c}{1+c}, \\ \dfrac{1}{1+c}\left(\dfrac{a}{(1+c)z - c + a}\right)^a, & \text{if } z > \dfrac{c}{1+c} \end{cases} \tag{2.4}$$

for some positive (a, c) values. Note that

1. $\phi_{\mathrm{u}}(z)$ is monotonically decreasing.
2. If c is fixed, then
$$\lim_{a\to\infty} \phi_{\mathrm{u}}(z) = \frac{1}{1+c}\exp\left(-(1+c)\left(z - \frac{c}{1+c}\right)\right).$$
Thus, if $c = 0$, then $\lim_{a\to\infty} \phi_{\mathrm{u}}(z) = \exp(-z)$ which is equal to the exponential loss function used in logistic regression.
3. If a is fixed, then $\lim_{c\to\infty} \phi_{\mathrm{u}}(z) = \max(1 - z, 0)$, since $\lim_{c\to\infty} \phi_{\mathrm{u}}(z)|_{z=c/(1+c)} = 0$ where this is the hinge loss function used in SVM.

See Figure 1(a) which shows loss functions for each (a, c) pair.

For the unified AUC maximizer, adapting ϕ_{u}, we can construct a ranking function by minimizing the following regularized empirical risk

$$\sum_{i\in I_+, j\in I_-} \phi_{\mathrm{u}}\left(-(f(\mathbf{x}_i) - f(\mathbf{x}_j))\right) + \lambda\|\beta\|_1, \tag{2.5}$$

where $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$ and $\lambda \ge 0$ is a regularization parameter controlling the complexity and sparsity of the ranking function.

Note that using $\phi_{\mathrm{u}}(z)$ in bipartite ranking problem can enable us to estimate $\Pr(Y > \tilde{Y}|X = \mathbf{x}, \tilde{X} = \tilde{\mathbf{x}})$ called the expected ranking accuracy presented in Lemma 1. Figure 1(b) depicts the corresponding transformed probabilities from the difference between two scores $\triangle f(\mathbf{x}, \tilde{\mathbf{x}}) = f(\mathbf{x}) - f(\tilde{\mathbf{x}}) = (\mathbf{x} - \tilde{\mathbf{x}})'\beta$ for various pair of (a, c) values respectively. The loss function of left figure on upper panel is close to $(1 - z)_+$ in SVM. The loss function of right figure on lower panel is close to $\exp(-z)$ or $\log(1 + \exp(-z))$ in logistic regression. Thus, for the larger c value, the estimates are closer to the solution of SVM. Also, for the larger a value, the estimates are closer to the solution of logistic regression. Moreover, it is possible to construct hybrids of loss functions used in SVM and logistic regression by choosing a and c. Especially, the hybrid method has a characteristic which maximizes large margin and estimates posterior simultaneously. For example, see the left figure on lower panel of Figure 1(b). The corresponding probability (of $(a, c) = (10, 1)$) is 1/2 in the center of the difference $f(\mathbf{x}) - f(\tilde{\mathbf{x}})$. That is, the hybrid method tries to maximize large margin as SVM in the center region. Except that region, it can provide posterior as logistic regression.

Lemma 1. *For given c and a values, let the expected ranking accuracy be* $\eta(\mathbf{x}, \tilde{\mathbf{x}}) = \Pr(Y > \tilde{Y}|X = \mathbf{x}, \tilde{X} = \tilde{\mathbf{x}})$ *then the corresponding posterior probability* $\eta(\mathbf{x}, \tilde{\mathbf{x}})$ *between difference of two scores* $\triangle f(\mathbf{x}, \tilde{\mathbf{x}})$ *is*

$$\eta(\mathbf{x}, \tilde{\mathbf{x}}) = \begin{cases} \dfrac{\gamma_1}{1+\gamma_1}, & \text{if } \triangle f(\mathbf{x}, \tilde{\mathbf{x}}) \le -\dfrac{c}{1+c}, \\ \dfrac{1}{2}, & \text{if } |\triangle f(\mathbf{x}, \tilde{\mathbf{x}})| < \dfrac{c}{1+c}, \\ \dfrac{1}{1+\gamma_2}, & \text{if } \triangle f(\mathbf{x}, \tilde{\mathbf{x}}) \ge \dfrac{c}{1+c}, \end{cases} \tag{2.6}$$

1. Initialize $\beta(0) = \mathbf{0}$
2. Repeat until $\delta_j(\eta) = 0, \forall j$
   (a) Compute $\{\delta_j(\eta),\ j = 1, \ldots, p\}$
   (b) Set $S = \{j : \delta_j(\eta) \cdot \beta_j(\eta)\} < 0$
   (c) Find updating coordinate:
   If S = empty, $j^* = \arg\max_j |\delta_j(\eta)|$, else $j^* = \arg\max_{j \in S} |\delta_j(\eta)|$
   (d) Update: $\beta_{j^*}(\eta + \triangle\eta) = \beta_{j^*} + \triangle\eta \text{sign}(\delta_{j^*}(\eta))$
   (e) Increase the regularization parameter: $\eta \longleftarrow \eta + \triangle\eta$

Figure 2: *Unified ranking algorithm using GPS (pseudo code)*

where

$$\gamma_1 = \left(\frac{a}{-(1+c)\triangle f(\mathbf{x}, \tilde{\mathbf{x}}) - c + a} \right)^{a+1} \quad \textit{and} \quad \gamma_2 = \left(\frac{a}{(1+c)\triangle f(\mathbf{x}, \tilde{\mathbf{x}}) - c + a} \right)^{a+1}.$$

In this paper, we consider the linear model $f(\mathbf{x}) = \beta_0 + \mathbf{x}'\beta$. Then the intercept β_0 is nuisance parameter since $f(\mathbf{x}_i) - f(\mathbf{x}_j) = \mathbf{x}'_{ij}\beta$ where $\mathbf{x}_{ij} = \mathbf{x}_i - \mathbf{x}_j$ in Equation (2.3). This means if we are interested in only ranks, then the intercept is not necessary. However, if we want to classify further, then the estimation of intercept is required. For this, let $\hat{h}(\mathbf{x}) = \mathbf{x}'\hat{\beta}$ and the order statistics of $\hat{h}(\mathbf{x})$ be $\hat{h}_{(k)}$, $k = 1, \ldots, n$ ($\hat{h}_{(1)} < \cdots < \hat{h}_{(n)}$). Then the natural estimate for β_0 is any value achieving minimum misclassification error rate among values such that $-\hat{h}_{(k)} < \hat{\beta}_0 < -\hat{h}_{(k-1)}$ ($e.g, \hat{\beta}_0 = -(\hat{h}_{(k-1)} + \hat{h}_{(k)})/2$).

2.3. Computation

In this section, given (a, c) values we note an optimization algorithm for Equation (2.5). Since n_+n_- pairs for the computation are used in the bipartite ranking problem, as either n_+ or n_- grows, the corresponding computational complexity increases much faster than that of the binary classification problem. Therefore, to scale up the problem, a type of entire solution path algorithm is necessary. The main advantage of the entire solution path algorithm is that it gives us all solutions where the complexity is equal to that of an usual estimator without a regularization. So, in determining the optimal λ, the entire solution path algorithm would be more efficient.

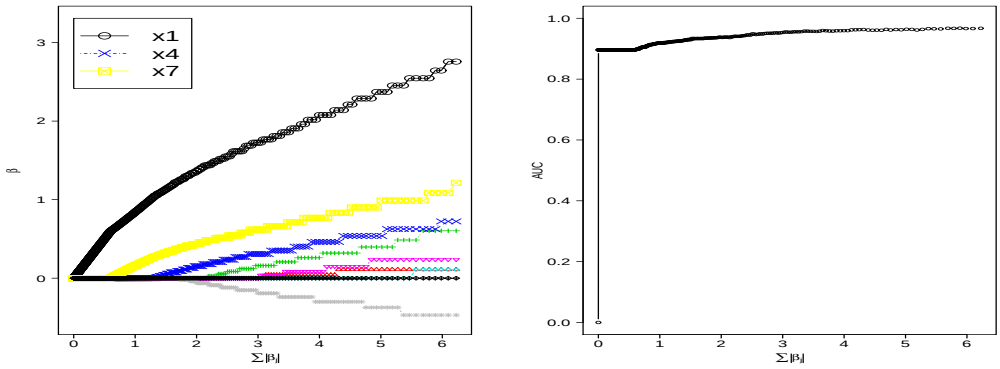
For this purpose, we are to apply the generalized path seeking algorithm(GPS) proposed by Friedman (2008). This algorithm can construct the first order of approximation of the entire solution path easily.

To apply GPS algorithm to Equation (2.5), we define some notations. Let the empirical risk be $R(\beta) = \sum_{i \in I_+, j \in I_-} \phi_{\text{u}}(-(x_i - x_j)'\beta)$ and $J(\beta) = \|\beta\|_1$. Let the working parameter of λ be η which measures length along the path and $\triangle\eta > 0$ be its a small increment. Let $\hat{\beta}(\eta)$ be solution at $\lambda = \eta$ and $\hat{\beta}(\eta + \triangle\eta)$ be solution at $\lambda = \eta + \triangle\eta$ similarly. The GPS algorithm finds the solution for each λ where increment size is $\triangle\eta$. It starts from $\lambda = 0$ and ends to $\lambda = \infty$. At current $\lambda = \eta$, it updates solution coodinatewisely and finds next solution at $\lambda = \eta + \triangle\eta$ iteratively.

Figure 2 summarizes the learning algorithm for the unified framework using GPS where we define gradients for the empirical risk, penalty and the ratio of two gradients:

$$g_j(\eta) = -\left[\frac{\partial R(\beta)}{\partial \beta_j} \right]_{\beta = \hat{\beta}(\eta)}, \quad p_j(\eta) = \left[\frac{\partial J(\beta)}{\partial |\beta_j|} \right]_{\beta = \hat{\beta}(\eta)} \text{ and } \delta_j(\eta) = \frac{g_j(\eta)}{p_j(\eta)},$$

for $j = 1, \ldots, p$.



(a) The approximated entire solution paths (b) Trajectory of AUC

Figure 3: *The solution paths of β and AUC value according to $\sum_{j=1}^{p}|\beta_j|$ for $(a, c) = (100, 0)$ from the simulated model.*

To show computational fitting process, we consider a simple model. The simulation model is $\log \Pr(Y = 1|\mathbf{x})/\Pr(Y = -1|X = \mathbf{x}) = \sum_{k=1}^{9}\beta_k x_k$ where $\mathbf{x} = (x_1, \ldots, x_9)'$ is a multivariate Gaussian random vector with mean 0 and covariates are mutually independent. We fix the sample size 100 and set the true $\beta = (3, 0, 0, 1.5, 0, 0, 2, 0, 0)'$. The variables x_1, x_4 and x_7 are signal inputs and others are noise inputs. Figure 3(a) draws the approximated entire solution paths of the estimates from the simulated model for $(a, c) = (100, 0)$. Figure 3(b) shows that the trajectory of AUC values according to $\sum_{j=1}^{p}|\beta_j|$ which implies that the proposed learning algorithm can successfully construct the solution paths.

3. Numerical Study

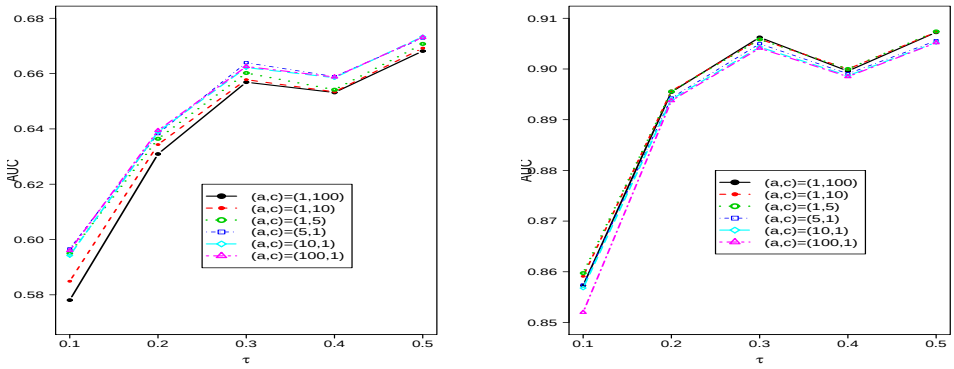
In this section, we investigate the finite sample performance of estimators of the proposed framework via simulation experiments. In particular, we compare the estimators in terms of predicted AUC values.

For simulation data, we generate a sample of size n as follows. Let μ^+ be a 100-dimensional vector whose first 3 entries are D and the other 97 entries are zero and let $\mu^- = -\mu^+$. Then, we generate $\mathbf{x}$ from $N_p(\mu^+, \Sigma)$ and assign $y = 1$ for the first $n \times \tau$ (τ represents asymmetric ratio of two populations) observations and generate $\mathbf{x}$ from $N_p(\mu^-, \Sigma)$ and assign $y = -1$ for the last $n - n \times \tau$ observations. We let the (k, l) entry of Σ be $0.3^{|k-l|}$. Note that all input variables except the first q are noisy.

We consider two scenarios to investigate characteristics of methods based on maximizing large margin and estimating posterior. The first scenario considers a large mean difference with $D = 0.5$. In the second scenario, mean difference is set to $D = 0.25$. In both scenarios, we investigate the performance with $(a, c) = (1, 100), (1, 10), (1, 5), (5, 1), (10, 1)$ and (100,1) varying asymmetric ratio $\tau = 0.1, 0.2, 0.3, 0.4$ and 0.5.

The results about AUC values are presented in Figure 4. The values are the averages based on 20 repetitions of the simulation. The regularization parameter is selected by the validation data set with size 200. The AUC values are measured on independent test samples of size 1000.

Figure 4 shows that the methods have better AUC for the larger asymmetric ratio. We have a conjecture that such the phenomena is due to the relation of AUC and misranking in Equation (2.1).



(a) Small mean difference with $D = 0.25$ (b) Large mean difference with $D = 0.5$

Figure 4: *Results for estimators (τ: asymmetric ratio of two populations).*

That is, it suggests that there is a necessity to adjust the cost of two populations like nonstandard binary classification problem (Kim, 2004).

Second, we can see that methods with more smoother loss function (with large a value) are better in AUC when the difference D is small, but vice versa when the difference D is large. Such results are parallel to the empirical studies in the classification problem of Kim *et al.* (2005). Since the methods based on maximizing large margin (*e.g,* SVM) estimate only the decision boundary, the large mean difference between two classes induces the better decision boundary. Meanwhile, in such the situation, the methods based on estimating posterior of whole area (*e.g,* logistic regression) would be inefficient (Bartlett and Tewari, 2007).

4. Discussion

In this paper, we set up the unified framework in the problem maximizing AUC criterion directly. However, from simulated data analysis, the performance of hybrid methods are not better than SVM and logistic regression. But, comparing the methods based on SVM and logistic regression, hybrids methods have some interesting applications. For example, let's think a situation which has observations which are hard to rank. That is, if the expected ranking probability $\eta(\mathbf{x}, \tilde{\mathbf{x}})$ is around 1/2, it would be better to take more advanced tests rather than to make a decision right away. In Lemma 1, if $|\triangle f(\mathbf{x}, \tilde{\mathbf{x}})|$ is less than $c/(1+c)$, we can know that the corresponding probability is equal to 1/2. It means two samples $\mathbf{x}$ and $\tilde{\mathbf{x}}$ have nearly same rank or tie. Also, in the view of Bradley-Terry model, we can interpret that such two samples (or two teams) with $\eta(\mathbf{x}, \tilde{\mathbf{x}}) = 1/2$ can have the chance to break even. So, the hybrid methods can be applied for the purpose. We leave this issue as a future work.

References

Agarwal, S., Graepel, T., Herbrich, R., Harpeled, S. and Roth, D. (2005). Generalization bounds for the area under the ROC curve, *Journal of Machine Learning Research,* **6**, 393–425.

Bach, F., Heckerman, D. and Horvitz, E. (2006). Considering cost asymmetry in learning classifiers, *Journal of Machine Learning Research,* **7**, 1713–1741.

Bartlett, P. and Tewari, A. (2007). Sparseness vs estimating conditional probabilities: Some asymptotic results, *Journal of Machine Learning Research,* **8**, 775–790.

Brefeld, U. and Scheffer, T. (2005). Auc maximizing support vector learning, In *Proceedings of the ICML. 2005 Workshop on ROC Analysis in Machine Learning*.

Clémencon, S., Lugosi, G. and Vayatis, N. (2006). From ranking to classification: A statistical view, *From Data and Information Analysis to Knowledge Engineering,* 214–221.

Clémencon, S., Lugosi, G. and Vayatis, N. (2008). Ranking and empirical minimization of U-statistics, *The Annals of Statistics,* **36**, 844–874.

Cortes, C. and Mohri, M. (2004). Auc optimization vs. error rate minimization, In Flach, F. et al. (Eds.), *In Advances in Neural Information Processing Systems,* **16**, MIT Press, Cambridge.

Cortes, C. and Vapnik, V. (1995). Support-vector networks, *Machine Learning,* **20**, 273–297.

Freund, Y., Iyer, R., Schapire, R. E. and Singer, Y. (2003). An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research,* **4**, 933–969.

Friedman, J. (2008). Fast sparse regression and classification, *Technical Report,* Stanford University.

Joachims, T. (2002). Optimizing search engines using clickthrough data, *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*.

Kim, J. (2004). ROC and cost graphs for general cost matrix where correct classifications incur non-zero costs, *Communications of the Korean Statistical Society,* **11**, 21–30.

Kim, Y., Kim, K. and Song, S. (2005). Comparison of boosting and SVM, *Journal of Korean Data & Information Science Society,* **16**, 999–1012.

Liu, Y. and Zhang, H. H. (2009). The large margin unified machines: A bridge between hard and soft classification. *The 1st Institute of Mathematical Statistics Asia Pacific Rim Meeting & 2009 Conference of the Korean Statistical Society*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society: Series B,* **58**, 267–288.

Received September 2009; Accepted November 2009