Semi-Supervised AUC Optimization without Guessing Labels of Unlabeled Data

Zheng Xie and Ming Li

AAAI, 2018

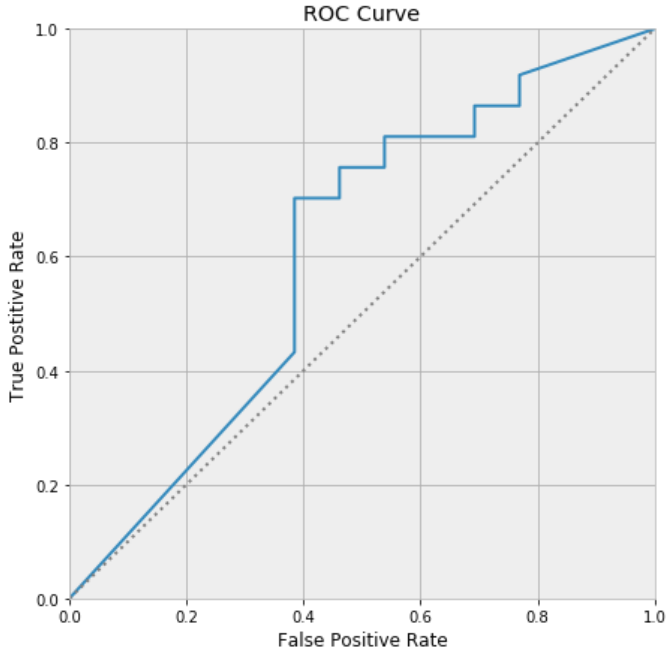
Introduction

- Semi-supervised Learning
  - Obtain the labels for the collected data is expensive
  - Semi-supervised learning are based on certain distributional assumption, e.g. cluster assumption and manifold assumption
  - Estimate the labels of unlabeled instances

- AUC Optimization for SSL
  - Rely on the **distributional assumptions**
  - Require **prior probability**
    - Sakai, Niu, and Sugiyama (2017) proposed an unbiased semi-supervised AUC optimization method based on positive-unlabeled learning

- Proposed Method
  - Do not depend on any distributional assumptions
  - Do not need to know prior probability

AUC



$$TPR = \frac{TP}{TP + FN} \longrightarrow P$$

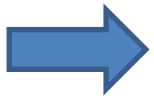
$$FPR = \frac{FP}{TN + FP} \longrightarrow N$$

The AUC is the probability the model will score a randomly chosen positive class higher than a randomly chosen negative class.

AUC Optimization

- Supervised learning

$$\mathcal{X}_{\mathrm{P}} := \{\boldsymbol{x}_i\}_{i=1}^{n_{\mathrm{P}}} \overset{\text{i.i.d.}}{\sim} p_{\mathrm{P}}(\boldsymbol{x}) := p(\boldsymbol{x} \mid y = +1)$$
$$\mathcal{X}_{\mathrm{N}} := \{\boldsymbol{x}'_j\}_{j=1}^{n_{\mathrm{N}}} \overset{\text{i.i.d.}}{\sim} p_{\mathrm{N}}(\boldsymbol{x}) := p(\boldsymbol{x} \mid y = -1)$$



$$f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$$

- Since AUC is equivalent to the probability of a randomly drawn positive instance being ranked before a randomly drawn negative instance, it can be formulated as

$$\mathrm{AUC} = 1 - \underbrace{\mathop{\mathbb{E}}_{\boldsymbol{x} \in \mathcal{X}_{\mathrm{P}}} \left[\mathop{\mathbb{E}}_{\boldsymbol{x}' \in \mathcal{X}_{\mathrm{N}}} \left[\ell_{01}(\boldsymbol{w}^\top (\boldsymbol{x} - \boldsymbol{x}')) \right] \right]}_{R_{\mathrm{PN}}}$$

AUC Optimization in SSL

- Rely on the distributional assumptions
  - These assumptions may be violated in many real-world applications

- Require an accurate estimation of the prior probability
  - It is usually difficult especially when the number of labeled data is extremely small.

We can achieve unbiased semi-supervised AUC optimization without distributional assumptions or prior knowledge about the distribution or class prior probabilities.

Theorem

In AUC optimization, $\min R_{PN} \Longleftrightarrow \min R_{PU}, \min R_{NU}$

$$R_{\mathrm{PU}} = \mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_{\mathrm{P}}}\left[\mathop{\mathbb{E}}_{\boldsymbol{x}''\in\mathcal{X}_{\mathrm{U}}}\left[\ell_{01}(\boldsymbol{w}^{\top}(\boldsymbol{x}-\boldsymbol{x}''))\right]\right],$$
$$R_{\mathrm{NU}} = \mathop{\mathbb{E}}_{\boldsymbol{x}''\in\mathcal{X}_{\mathrm{U}}}\left[\mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_{\mathrm{N}}}\left[\ell_{01}(\boldsymbol{w}^{\top}(\boldsymbol{x}''-\boldsymbol{x}'))\right]\right].$$

Proof:

$$\mathcal{X}_{\mathrm{U}} := \{\boldsymbol{x}''_k\}_{k=1}^{n_{\mathrm{U}}} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}) = \theta_P p_{\mathrm{P}}(\boldsymbol{x}) + \theta_N p_{\mathrm{N}}(\boldsymbol{x})$$

$$R_{\mathrm{PU}} = \mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_{\mathrm{P}}}\left[\mathop{\mathbb{E}}_{\boldsymbol{x}''\in\mathcal{X}_{\mathrm{U}}}\left[\ell_{01}(\boldsymbol{w}^{\top}(\boldsymbol{x}-\boldsymbol{x}''))\right]\right]$$
$$= \mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_{\mathrm{P}}}\left[\theta_P \mathop{\mathbb{E}}_{\bar{\boldsymbol{x}}\in\mathcal{X}_{\mathrm{P}}}\left[\ell_{01}(\boldsymbol{w}^{\top}(\boldsymbol{x}-\bar{\boldsymbol{x}}))\right] + \theta_N \mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_{\mathrm{N}}}\left[\ell_{01}(\boldsymbol{w}^{\top}(\boldsymbol{x}-\boldsymbol{x}'))\right]\right]$$
$$= \frac{1}{2}\theta_P + \theta_N \mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{X}_{\mathrm{P}}}\left[\mathop{\mathbb{E}}_{\boldsymbol{x}'\in\mathcal{X}_{\mathrm{N}}}\left[\ell_{01}(\boldsymbol{w}^{\top}(\boldsymbol{x}-\boldsymbol{x}'))\right]\right]$$

$$\frac{1}{2}\left[l_{01}(w(x_1-x_2)) + l_{01}(w(x_2-x_1))\right] = \frac{1}{2}$$

Unbiased Estimation

$$\theta_P + \theta_N = 1$$

$$R_{\mathrm{PU}} = \theta_N R_{\mathrm{PN}} + \frac{1}{2}\theta_P$$

$$R_{\mathrm{NU}} = \theta_P R_{\mathrm{PN}} + \frac{1}{2}\theta_N$$

$$\boxed{R_{\mathrm{PU}} + R_{\mathrm{NU}} - \frac{1}{2} = R_{\mathrm{PN}}}$$

Conduct unbiased AUC risk estimation without knowing the class prior probabilities θ_P and θ_N.

Alg-1. SAMULT

Semi-supervised **A**UC **M**aximization by treating the **UnL**abeled data in **T**wo ways

Objective: $\min_{\boldsymbol{w}} \quad \gamma \widehat{R}_{\mathrm{PN}} + (1-\gamma)\left(\widehat{R}_{\mathrm{PU}} + \widehat{R}_{\mathrm{NU}} - \frac{1}{2}\right) + \lambda ||\boldsymbol{w}||^2$

supervised AUC risk

unbiased semi-supervised AUC risk

Where,

$$\widehat{R}_{\mathrm{PN}} = \frac{1}{n_{\mathrm{P}} n_{\mathrm{N}}} \sum_{\boldsymbol{x} \in \mathcal{X}_{\mathrm{P}}} \sum_{\boldsymbol{x}' \in \mathcal{X}_{\mathrm{N}}} \ell(\boldsymbol{w}^\top (\boldsymbol{x} - \boldsymbol{x}'))$$

$$\widehat{R}_{\mathrm{PU}} = \frac{1}{n_{\mathrm{P}} n_{\mathrm{U}}} \sum_{\boldsymbol{x} \in \mathcal{X}_{\mathrm{P}}} \sum_{\boldsymbol{x}'' \in \mathcal{X}_{\mathrm{U}}} \ell(\boldsymbol{w}^\top (\boldsymbol{x} - \boldsymbol{x}''))$$

$$\widehat{R}_{\mathrm{NU}} = \frac{1}{n_{\mathrm{U}} n_{\mathrm{N}}} \sum_{\boldsymbol{x}'' \in \mathcal{X}_{\mathrm{U}}} \sum_{\boldsymbol{x}' \in \mathcal{X}_{\mathrm{N}}} \ell(\boldsymbol{w}^\top (\boldsymbol{x}'' - \boldsymbol{x}'))$$

$$\gamma \in [0, 1]$$

$$\lambda \geq 0$$

SAMULT does not determine the decision boundary.

Analytical Solution

$$\widehat{\boldsymbol{w}} = (\gamma \boldsymbol{H}_{\mathrm{PN}} + (1-\gamma)(\boldsymbol{H}_{\mathrm{PU}} + \boldsymbol{H}_{\mathrm{NU}}) + \lambda \boldsymbol{I}_d)^{-1} \\ (\gamma \boldsymbol{h}_{\mathrm{PN}} + (1-\gamma)(\boldsymbol{h}_{\mathrm{PU}} + \boldsymbol{h}_{\mathrm{NU}}))\ ,$$

where

$$\begin{aligned}
\boldsymbol{h}_{\mathrm{PN}} =& \frac{1}{n_{\mathrm{P}}} \boldsymbol{X}_{\mathrm{P}}^{\top} \mathbf{1}_{n_{\mathrm{P}}} - \frac{1}{n_{\mathrm{N}}} \boldsymbol{X}_{\mathrm{N}}^{\top} \mathbf{1}_{n_{\mathrm{N}}}\ , \\
\boldsymbol{h}_{\mathrm{PU}} =& \frac{1}{n_{\mathrm{P}}} \boldsymbol{X}_{\mathrm{P}}^{\top} \mathbf{1}_{n_{\mathrm{P}}} - \frac{1}{n_{\mathrm{U}}} \boldsymbol{X}_{\mathrm{U}}^{\top} \mathbf{1}_{n_{\mathrm{U}}}\ , \\
\boldsymbol{h}_{\mathrm{NU}} =& \frac{1}{n_{\mathrm{U}}} \boldsymbol{X}_{\mathrm{U}}^{\top} \mathbf{1}_{n_{\mathrm{U}}} - \frac{1}{n_{\mathrm{N}}} \boldsymbol{X}_{\mathrm{N}}^{\top} \mathbf{1}_{n_{\mathrm{N}}}\ , \\
\boldsymbol{H}_{\mathrm{PN}} =& \frac{1}{n_{\mathrm{P}}} \boldsymbol{X}_{\mathrm{P}}^{\top} \boldsymbol{X}_{\mathrm{P}} - \frac{1}{n_{\mathrm{P}} n_{\mathrm{N}}} \boldsymbol{X}_{\mathrm{P}}^{\top} \mathbf{1}_{n_{\mathrm{P}}} \mathbf{1}_{n_{\mathrm{N}}}^{\top} \boldsymbol{X}_{\mathrm{N}} \\
& - \frac{1}{n_{\mathrm{P}} n_{\mathrm{N}}} \boldsymbol{X}_{\mathrm{N}}^{\top} \mathbf{1}_{n_{\mathrm{N}}} \mathbf{1}_{n_{\mathrm{P}}}^{\top} \boldsymbol{X}_{\mathrm{P}} + \frac{1}{n_{\mathrm{N}}} \boldsymbol{X}_{\mathrm{N}}^{\top} \boldsymbol{X}_{\mathrm{N}}\ , \\
\boldsymbol{H}_{\mathrm{PU}} =& \frac{1}{n_{\mathrm{P}}} \boldsymbol{X}_{\mathrm{P}}^{\top} \boldsymbol{X}_{\mathrm{P}} - \frac{1}{n_{\mathrm{P}} n_{\mathrm{U}}} \boldsymbol{X}_{\mathrm{P}}^{\top} \mathbf{1}_{n_{\mathrm{P}}} \mathbf{1}_{n_{\mathrm{U}}}^{\top} \boldsymbol{X}_{\mathrm{U}} \\
& - \frac{1}{n_{\mathrm{P}} n_{\mathrm{U}}} \boldsymbol{X}_{\mathrm{U}}^{\top} \mathbf{1}_{n_{\mathrm{U}}} \mathbf{1}_{n_{\mathrm{P}}}^{\top} \boldsymbol{X}_{\mathrm{P}} + \frac{1}{n_{\mathrm{U}}} \boldsymbol{X}_{\mathrm{U}}^{\top} \boldsymbol{X}_{\mathrm{U}}\ , \\
\boldsymbol{H}_{\mathrm{NU}} =& \frac{1}{n_{\mathrm{U}}} \boldsymbol{X}_{\mathrm{U}}^{\top} \boldsymbol{X}_{\mathrm{U}} - \frac{1}{n_{\mathrm{U}} n_{\mathrm{N}}} \boldsymbol{X}_{\mathrm{U}}^{\top} \mathbf{1}_{n_{\mathrm{U}}} \mathbf{1}_{n_{\mathrm{N}}}^{\top} \boldsymbol{X}_{\mathrm{N}} \\
& - \frac{1}{n_{\mathrm{U}} n_{\mathrm{N}}} \boldsymbol{X}_{\mathrm{N}}^{\top} \mathbf{1}_{n_{\mathrm{N}}} \mathbf{1}_{n_{\mathrm{U}}}^{\top} \boldsymbol{X}_{\mathrm{U}} + \frac{1}{n_{\mathrm{N}}} \boldsymbol{X}_{\mathrm{N}}^{\top} \boldsymbol{X}_{\mathrm{N}}\ ,
\end{aligned}$$

SAMULT^{P+U}

When only the positive and unlabeled data are available

$$\min_{\boldsymbol{w}} \quad \gamma\widehat{R}_{\mathrm{PN}} + (1-\gamma)(\widehat{R}_{\mathrm{PU}} + \widehat{R}_{\mathrm{NU}} - \frac{1}{2}) + \lambda||\boldsymbol{w}||^2$$

$$\min_{\boldsymbol{w}} \quad \widehat{R}_{\mathrm{PU}}(\boldsymbol{w}) + \lambda||\boldsymbol{w}||^2$$

Alg-2. SAMPURA

Semi-supervised **A**UC **M**aximization by **P**artitioning **U**nlabeled data at **RA**ndom

$$\mathcal{X}_{\mathrm{U}} \Rightarrow \begin{matrix} \mathcal{X}_{\mathrm{U}+} \\ \mathcal{X}_{\mathrm{U}-} \end{matrix} \Rightarrow \begin{matrix} \mathcal{X}_{\mathrm{P}'} = \mathcal{X}_{\mathrm{P}} \cup \mathcal{X}_{\mathrm{U}+} \\ \mathcal{X}_{\mathrm{N}'} = \mathcal{X}_{\mathrm{N}} \cup \mathcal{X}_{\mathrm{U}-} \end{matrix}$$

$\mathcal{X}_{\mathrm{P}'}$ should be ranked before $\mathcal{X}_{\mathrm{N}}$

$\mathcal{X}_{\mathrm{P}}$ should be ranked before $\mathcal{X}_{\mathrm{N}'}$

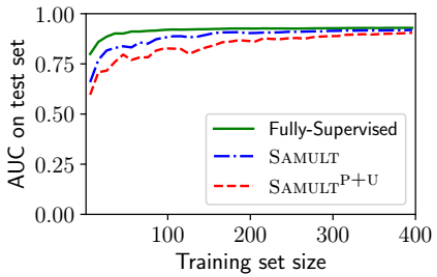
$$\min_{\boldsymbol{w}} \quad \widehat{R}_{\mathrm{P}'\mathrm{N}} + \widehat{R}_{\mathrm{P}\mathrm{N}'} + \lambda ||\boldsymbol{w}||^2$$

take the average of those ***w***s to construct the final ensemble

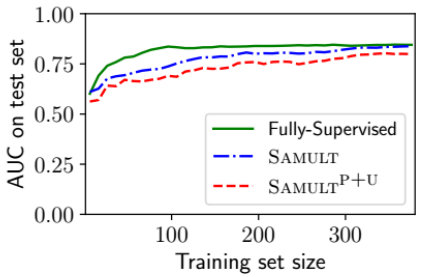
Experiment - 1

The model that minimizes the PU-AUC risk or the PNU-AUC risk converges to the supervised case

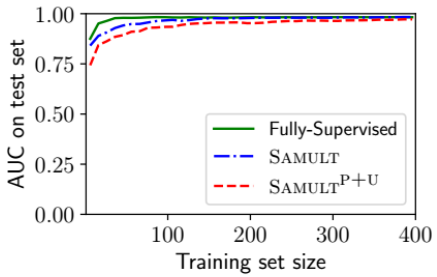
AUC



(a) *australian*



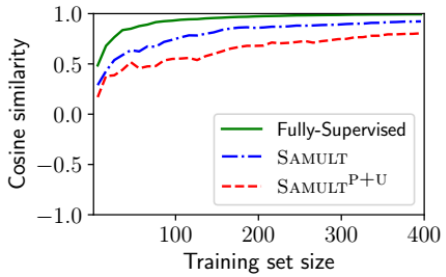
(b) *clean1*



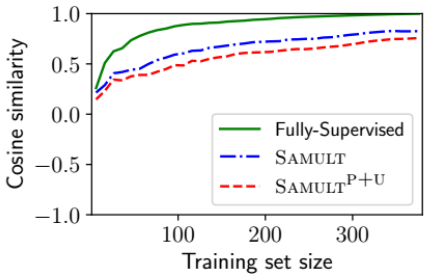
(c) *wdbc*

10% data in training set is labeled

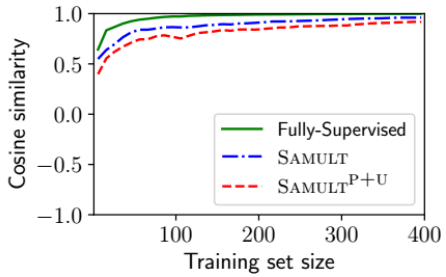
$\cos\langle \widehat{\boldsymbol{w}}, \widehat{\boldsymbol{w}}^* \rangle$



(a) *australian*



(b) *clean1*



(c) *wdbc*

$\widehat{\boldsymbol{w}}^*$ is learned from all available data with label

Experiment - 2

Compare our methods to the state-of-the-art semi-supervised AUC optimization methods.

- **SAMULT**
- **SAMPURA**
- **SSRankBoost** (Amini, Truong, and Goutte 2008), a boosting based algorithm for learning a bipartite ranking function with partially labeled data
- **PNU-AUC** (Sakai, Niu, and Sugiyama 2017), which is a semi-supervised AUC optimization method based on positive-unlabeled learning.
- **Supervised AUC optimization**
- **Logistic regression**

Experiment - 2

Dataset	Supervised	Log. Reg.	SSRankBoost	PNU-AUC	SAMULT	SAMPURA
australian	.879±.029	.860±.027	.886±.013	**.903±.009**	**.903±.009**	**.903±.009**
breast	.655±.097	.625±.095	.647±.065	**.701±.029**	**.701±.029**	**.704±.026**
breastw	.987±.009	.980±.006	.984±.013	.992±.001	**.996±.001**	**.996±.001**
clean1	.760±.062	.725±.060	.737±.050	.767±.042	.777±.039	**.782±.038**
colic	.829±.112	.818±.074	.721±.062	.858±.013	**.869±.013**	**.870±.013**
colic.orig	.647±.093	.645±.076	.612±.081	.644±.048	**.658±.049**	**.663±.044**
credit-a	.893±.024	.886±.023	.885±.013	**.906±.008**	**.906±.007**	**.906±.008**
credit-g	.719±.043	.709±.030	.665±.027	.748±.018	.748±.018	**.761±.017**
fourclass	.825±.023	.826±.026	.692±.029	**.827±.008**	**.827±.008**	**.828±.006**
german	.683±.057	.672±.048	.709±.025	**.727±.019**	**.727±.019**	**.729±.017**
haberman	.551±.086	.530±.075	**.582±.067**	.547±.051	.551±.045	.556±.043
heart	.857±.065	.842±.060	.823±.042	**.876±.025**	**.876±.025**	**.878±.024**
house	**.975±.038**	.961±.015	**.972±.034**	**.979±.012**	**.979±.012**	**.979±.011**
ijcnn1	**.912±.003**	.901±.004	.902±.002	.904±.009	**.913±.005**	**.915±.004**
madelon	.510±.037	.541±.020	**.571±.023**	.528±.029	.517±.027	.530±.022
parkinsons	.848±.129	.826±.082	.799±.051	**.860±.023**	**.860±.023**	**.863±.011**
phishing	.975±.097	.972±.001	.983±.003	.974±.002	.976±.002	**.985±.002**
vehicle	.932±.038	.922±.022	.912±.039	**.965±.020**	**.965±.020**	**.970±.014**
vote	.965±.038	.951±.015	**.972±.034**	**.979±.012**	**.979±.012**	**.979±.011**
wdbc	.971±.014	.963±.006	.964±.016	**.983±.006**	**.983±.006**	**.983±.005**
# Best/Comp.	2	0	4	11	15	18

Experiment - 3

Dataset	PU-RSVM	PU-AUC	SAMULT^{P+U}
australian	.844±.034	.898±.021	**.900±.019**
breast	.615±.104	**.701±.077**	**.701±.077**
breastw	.987±.009	.993±.002	**.996±.002**
clean1	.709±.072	.786±.050	**.796±.050**
colic	.807±.103	**.877±.060**	**.877±.060**
colic.orig	.621±.078	.650±.068	**.670±.061**
credit-a	.876±.028	**.912±.015**	**.912±.015**
credit-g	.688±.047	.755±.028	**.757±.027**
fourclass	.823±.031	**.832±.026**	**.832±.025**
german	.642±.045	.734±.034	**.736±.034**
haberman	**.572±.083**	.561±.081	.555±.080
heart	.835±.073	**.883±.039**	**.883±.040**
house	.945±.029	.980±.012	**.983±.011**
ijcnn1	**.927±.005**	.900±.011	.905±.012
madelon	.470±.015	**.533±.031**	.514±.032
parkinsons	.797±.089	**.870±.033**	**.870±.032**
phishing	.960±.008	.966±.005	**.970±.005**
vehicle	.942±.034	.959±.030	**.966±.025**
vote	.945±.029	.980±.012	**.983±.011**
wdbc	.967±.021	.984±.007	**.985±.007**
#Best/Com.	2	7	17

Degeneration to AUC Optimization for Positive and Unlabeled Data

PU-RSVM (Sellamanickam, Garg, and Selvaraj 2011):
A ranking SVM based method for positive and unlabeled data

PU-AUC (Sakai, Niu, and Sugiyama 2017):
A positive unlabeled AUC optimization method by optimizing an unbiased AUC risk estimator relies only on positive and unlabeled data

Do not need to know the prior.

Conclusion

- In semi-supervised AUC optimization, it is unnecessary to design sophisticated strategies to estimate the possible labels of the unlabeled data or the class prior probabilities.

- Proposed two semi-supervised AUC optimization methods: SAMULT and SAMPURA

- The positive-unlabeled AUC optimization problem can be addressed by a degenerated version of proposed method that simply treats the unlabeled data as negative.

Reference

- Zheng Xie, Ming Li. **Semi-Supervised AUC Optimization without Guessing Labels of Unlabeled Data**. AAAI, 2018.

- Sundararajan Sellamanickam, Priyanka Garg, Sathiya Keerthi Selvaraj. **A pairwise ranking based approach to learning with positive and unlabeled examples.** CIKM, 2011.

- Tomoya Sakai, Gang Niu, and Masashi Sugiyama. **Semi-Supervised AUC Optimization Based on Positive-Unlabeled Learning.** *Machine Learning*, 2017.

- Massih-Reza Amini, Tuong-Vinh Truong, Cyril Goutte. **A Boosting Algorithm for Learning Bipartite Ranking Functions with Partially Labeled Data.** SIGIR, 2008.