Active Sampling of Pairs and Points for Large-scale Linear Bipartite Ranking

Wei-Yuan Shen and Hsuan-Tien Lin

Department of Computer Science
& Information Engineering

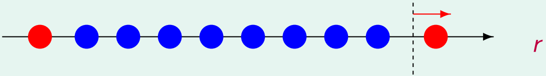
National Taiwan University

Short Talk for ACML @ Canberra, Australia
November 15, 2013

Introduction

Bipartite Ranking Problem

- **Input:** N examples $\mathbf{x}_n$ with $+1$/-1 labels y_n
- **Output:** ranking function $r(\mathbf{x})$ outputting real-values
- **Goal:** order positive instance higher than negative one as much as possible—equivalent to maximizing AUC(r)
- **Applications:** Information Retrieval, Bioinformatics, etc.
- **Related Problems:**
    - General Ranking: similar goal, different input
    - Binary Classification: same input, different goal



General Approaches

- **Pair-wise Approach**:
  - instance $\mathbf{x}$ of higher rank than $\mathbf{x}'$? $\rightarrow$ learn from *pairs*
  - Pros: consistent with learning goal, promising result
  - Cons: $O(N^2)$ number of pairs
- **Point-wise Approach**:
  - instance $\mathbf{x}$ positive? $\rightarrow$ learn from *points*
  - Pros: $\Theta(N)$ number of points
  - Cons: sometimes inferior performance

Active Sampling under Combined Ranking and Classification

- Active Sampling (AS): maintain efficiency
- Combined Ranking and Classification (CRC): enhance performance

—with linear SVM for efficiency

Baseline Work

Pair-wise SVM (RankSVM)

- $\mathcal{D}_{pair} = \left\{ \left(\mathbf{x}_{ij} = \mathbf{x}_i - \mathbf{x}_j, y_{ij} = sign(y_i - y_j) \right) : y_i \neq y_j \right\}$
- no-bias SVM on $\mathcal{D}_{pair}$ (Herbrich, 2000)

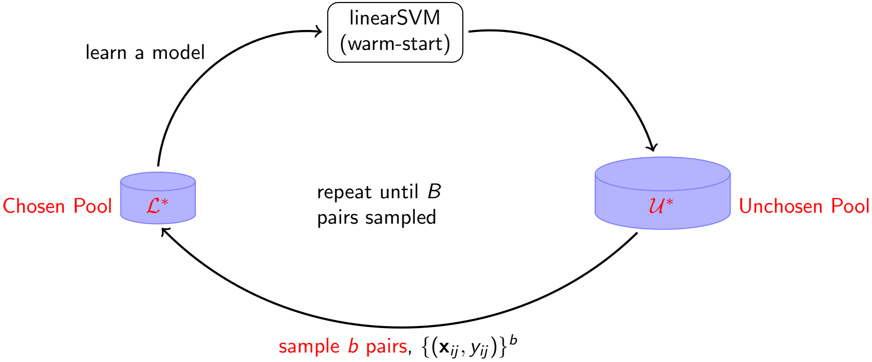
$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{\mathbf{x}_{ij} \in \mathcal{D}_{pair}} C_{ij} \cdot \text{hinge}(y_{ij}\mathbf{w}^T\mathbf{x}_{ij}) \ , \qquad (1)$$

- naïve RankSVM: $O(N^2)$ per iteration
- efficient RankSVM (Joachims, 2006): $O(N \log N)$ per iteration

Active Sampling

Motivation

- even $O(N \log N)$ can be infeasible when large-scale
- want: focus on a smaller (size-B) set of key pairs



Active Sampling: mimics Active Learning while 'knowing' true labels in $\mathcal{U}^*$

Sampling Strategies

given an unchosen pair $(\mathbf{x}_{ij}, y_{ij})$ & current ranking function $\mathbf{w}$:

- *closeness*$(\mathbf{x}_{ij}, y_{ij}, \mathbf{w}) = |y_{ij}\mathbf{w}^T\mathbf{x}_{ij}|$: how close to **w** the pair is ?
- *correctness*$(\mathbf{x}_{ij}, y_{ij}, \mathbf{w}) = -\text{hinge}(y_{ij}\mathbf{w}^T\mathbf{x}_{ij})$: how accurate the pair is ?

Hard Version Sampling

- select the pair with smallest *closeness* → *uncertainty sampling*
- select the pair with smallest *correctness* → *expected error reduction*
- finding lowest *closeness* or *correctness* pair: time consuming
- solution: **soft version** active sampling by rejection sampling

Soft Active Sampling Helps?

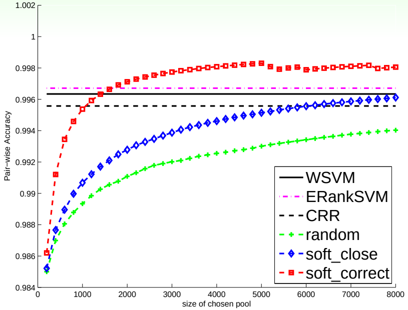


Figure: Performance Curves on **url**

- soft-correct > random: key pairs better than random
- soft-correct > ERankSVM: key pairs better than state-of-art
- soft-correct performs the best in general

Soft Active Sampling Helps?

soft-correct versus others based on t-test results (95% confidence level):

Data	WSVM	ERankSVM	CRR	Random	Soft-Close
letter	◯	△	△	△	△
protein	×	×	×	△	△
news20	◯	◯	◯	◯	◯
rcv1	◯	◯	◯	◯	◯
a9a	△	×	◯	△	×
bank	◯	◯	◯	◯	◯
ijcnn1	◯	◯	◯	◯	◯
shuttle	◯	◯	◯	◯	◯
mnist	×	×	◯	◯	×
connect	△	×	◯	◯	△
acoustic	◯	◯	◯	◯	◯
real-sim	◯	◯	◯	◯	◯
covtype	◯	◯	◯	◯	◯
url	◯	◯	◯	◯	◯
Total(win/loss/tie)	10 / 2 / 2	9 / 4 / 1	12 / 1 / 1	11 / 0 / 3	9 / 2 / 3

◯:win ×:loss △:tie

Combined Ranking and Classification

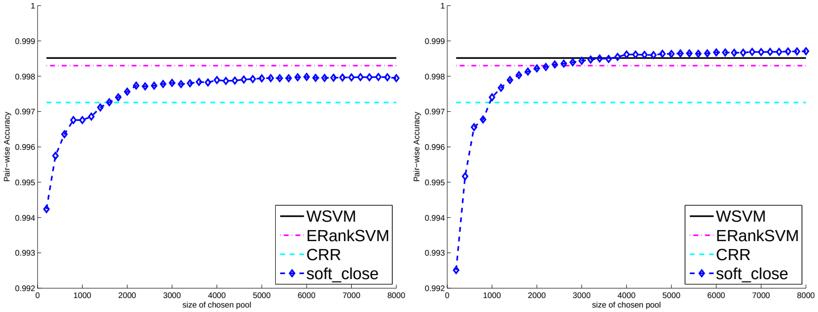
Motivation

- point-wise SVM is strong
- some points may also be valuable

Combined Ranking and Classification (CRC)

- unifies points and pairs for better performance
- *pair-wise* loss function: $\text{hinge}(y_{ij}\mathbf{w}^T(\mathbf{x}_i - \mathbf{x}_j))$
- *point-wise* loss function: $\text{hinge}(y_i\mathbf{w}^T\mathbf{x}_i) = \text{hinge}(y_i\mathbf{w}^T(\mathbf{x}_i - \mathbf{0}))$
- *point* and zero vector (opposite label) $\rightarrow$ *pseudo-pair*

CRC Helps?



(a) mnist (no pseudo)   (b) mnist (some pseudo)

- flexibility of CRC can be useful

Conclusion

Active Sampling under Combined Ranking and Classification

- AS successful in selecting key pairs
- CRC unifies *point-wise* and *pair-wise*
- experiments on 14 real-world large-scale data sets show promising performance, robustness and efficiency

Thank you! Any Questions?