Multi-label Active Learning with Auxiliary Learner

Chen-Wei Hung and **Hsuan-Tien Lin**

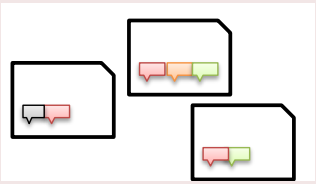
National Taiwan University

November 15, 2011

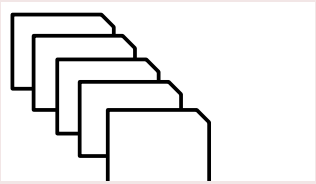
Multi-label Active Learning

multi-label data set:

- labeled pool $D_l = \{(\mathbf{x}'_n, \mathbf{y}'_n)\}$



- unlabeled pool $D_u = \{\mathbf{x}_n\}$



learning:

- train with data set to get decision functions $f_k(\mathbf{x})$

active:

- query labels of size-S set $D_s^* \subset D_u$
- move D_s^* & labels from D_u to D_l

- expensiveness of labeling, especially for multi-label
- active learning: allow "asking questions" (query labels)

hope: reduce labeling cost while maintaining good performance by **asking key questions**

Problem Setup

Given

- K-class problem with labeled pool D_l that contains (input $\mathbf{x}'_n$, label-set y'_n); y'_n expressed by $\{-1, +1\}^K$
- an unlabeled pool $D_u = \{\mathbf{x}_n\}$

Goal

a multi-label active learning algorithm that iteratively

- learn a **decent classifier** $f_k(\mathbf{x}) \in \mathbb{R}^K$ from D_l, with $\text{sign}(f_k(\mathbf{x}))$ used to predict k-th label
- choose a **key subset** D_s^* from D_u to be queried

and **improve performance of f_k efficiently** w.r.t. # queries

multi-label active learning:
newer and less-studied (than binary active learning)

Max. Loss Reduction with Max. Confidence (MMC)

State-of-the-art in Multi-label Active Learning

MMC: proposed by Yang et al.,
Effective Multi-label Active Learning for Text Classification, KDD, 2009

first-level learning:
get $g_k(\mathbf{x})$ by binary relevance SVM (BRSVM) from D_l

→

second-level learning:
get $f_k(\mathbf{x})$ by stacked logistic reg. (SLR) from D_l & $g_k(\mathbf{x})$

query: by maximum margin reduction using f_k and g_k

- binary relevance SVM (BRSVM): one binary SVM per label
- **promising practical performance**
  with some theoretical rationale

Motivation: How to improve MMC?

Multi-label Active Learning with Auxiliary Learner

Idea

digest the **essence** of MMC, and then extend for **improvement**

auxiliary learning:
get $g_k(\mathbf{x})$ by some $\mathcal{G}$ from D_l

major learning:
get $f_k(\mathbf{x})$ by some $\mathcal{F}$ from D_l

query by **disagreement of** g_k & f_k

- proposed framework: query with **two** learners—major & auxiliary
- major (original f_k):
  for **accurate predictions** of multi-label learning
- auxiliary:
  a **different** one to help **query decisions**

MMC
= (major: SLR) + (auxiliary: BRSVM) + (criterion: MMR)

Maximum Margin Reduction (MMR), Used by MMC

Intuition: Query by Version Space Reduction

$$\text{query set } D_s^* = \underset{|D_s|=S, D_s \subset D_u}{\text{argmax}} \left\{ V(\mathcal{G}, D_l) - V(\mathcal{G}, D_l \cup \text{labeled } D_s) \right\}$$

- V: size of version space (set of classifiers consistent to data)
- rationale: **smaller V $\rightarrow$ less ambiguity in learning $\rightarrow$ better**
- MMR: with some other assumptions[‡]

$$D_s^* \approx \text{top } S \text{ instances} \in D_u, \text{ ordered by } \sum_{k=1}^{K} \frac{1 - \text{sign}(f_k(\mathbf{x})) \cdot g_k(\mathbf{x})}{2}$$

equivalent MMR criterion: $-\sum_{k=1}^{K} \text{sign}(f_k(\mathbf{x})) \cdot g_k(\mathbf{x})$

[‡] Yang et al.,Effective Multi-label Active Learning for Text Classification, KDD09

Maximum Hamming Loss Reduction (HLR)

Intuition: Query by Hamming Loss Reduction

$$\text{query set } D_s^* = \underset{|D_s|=S, D_s \subset D_u}{\text{argmax}} \left\{ \text{HL}(\mathcal{G}, D_l) - \text{HL}(\mathcal{G}, D_l \cup \text{labeled } D_s) \right\}$$

- HL: Hamming loss made by learner $\mathcal{G}$
- rationale: **smaller HL $\rightarrow$ better performance in learning**
- HLR (our proposed criterion): with some assumptions

$$D_s^* \approx \text{top } S \text{ instances} \in D_u,$$
$$\text{ordered by } \sum_{k=1}^{K} \left[\!\left[\text{sign}(f_k(\mathbf{x})) \neq \text{sign}(g_k(\mathbf{x})) \right]\!\right]$$

equivalent HLR criterion:
$-\sum_{k=1}^{K} \text{sign}(f_k(\mathbf{x})) \cdot \text{sign}(g_k(\mathbf{x}))$

Quick Comparison between MMR and HLR

MMR

$$-\sum_{k=1}^{K} \text{sign}(f_k(\mathbf{x})) \cdot g_k(\mathbf{x})$$

- rationale: reduce V rapidly
- magnitude-sensitive  : few large g_k that disagree with $f_k \Longrightarrow$ **must query** (not robust to outliers)

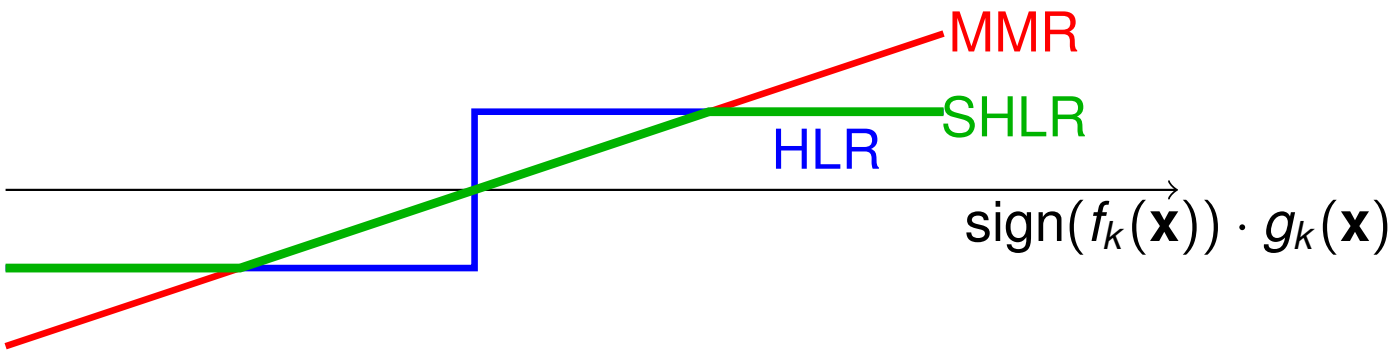
HLR

$$-\sum_{k=1}^{K} \text{sign}(f_k(\mathbf{x})) \cdot \text{sign}(g_k(\mathbf{x}))$$

- rationale: reduce HL rapidly
- magnitude-insensitive  : useful ambiguity information in g_k **lost** (not aware of details)

better criterion by combining the two? Yes!

Soft Hamming Loss Reduction



- rationale:
  $g_k(\mathbf{x})$ large—HLR to be robust to magnitude
  $g_k(\mathbf{x})$ small—MMR to keep ambiguity information
- Soft HLR:

$$D_s^* = \text{top } S \text{ instances} \in D_u,$$
$$\text{ordered by } \sum_{k=1}^{K} \text{clip}\Big(-\text{sign}(f_k(\mathbf{x})) \cdot g_k(\mathbf{x}), -1, 1\Big)$$

which is better? SHLR, HLR or MMR?

Experiment

Query Criteria

- Random: use neither auxiliary nor major
- BinMin: use only auxiliary but not major
- SHLR, HLR, MMR: use both auxiliary & major

Setting (same as used by Yang et al. to evaluate MMC)

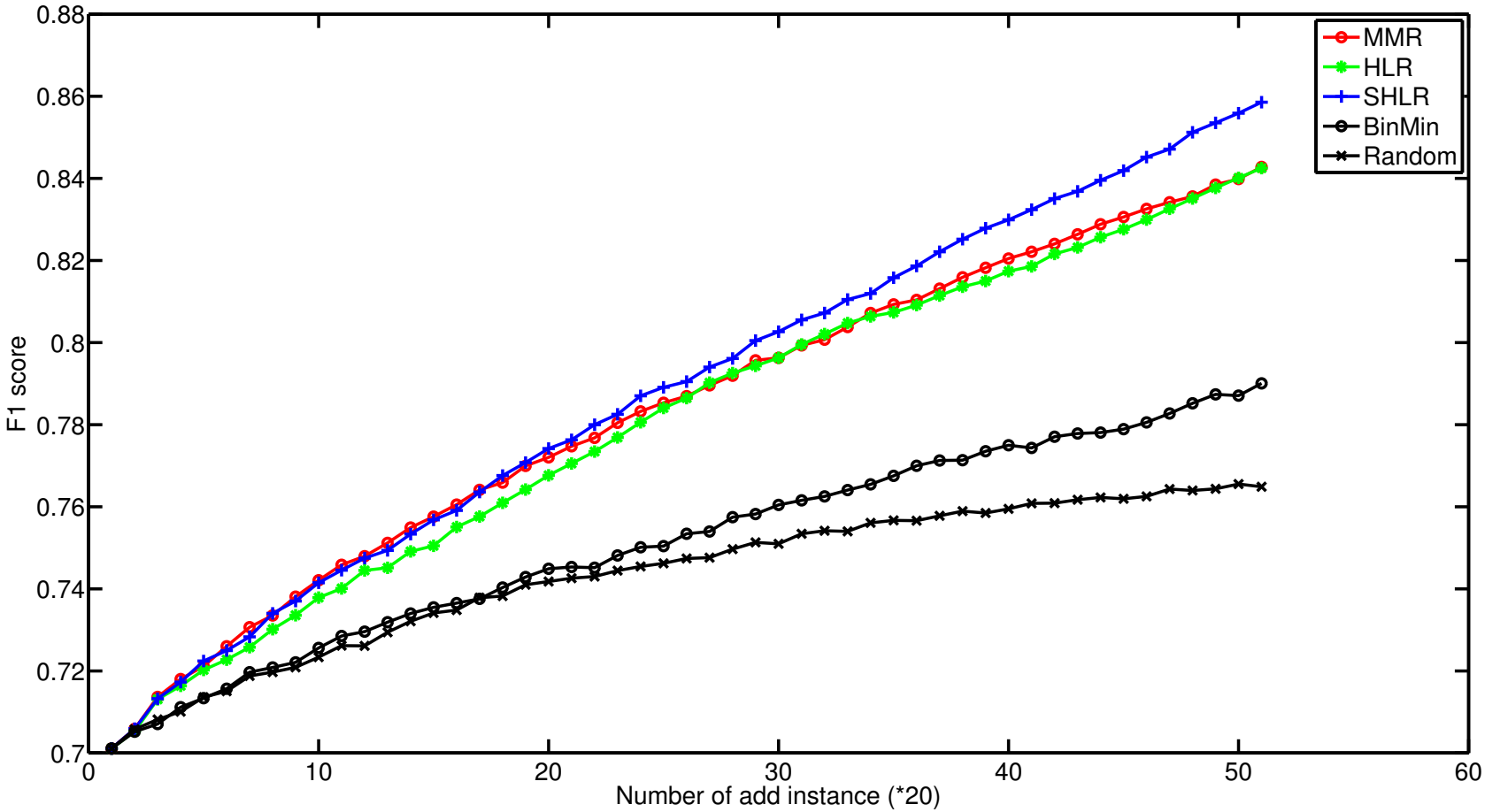
- D_l size: initial 500 to final 1500, step by $S = 20$

Major/Auxiliary Combination

- **major = SLR[BRSVM]; auxiliary = BR(SVM): used by MMC**
- major = CC(SVM); auxiliary = BR(SVM)
- major = SLR[BRSVM]; auxiliary = CC(SVM)

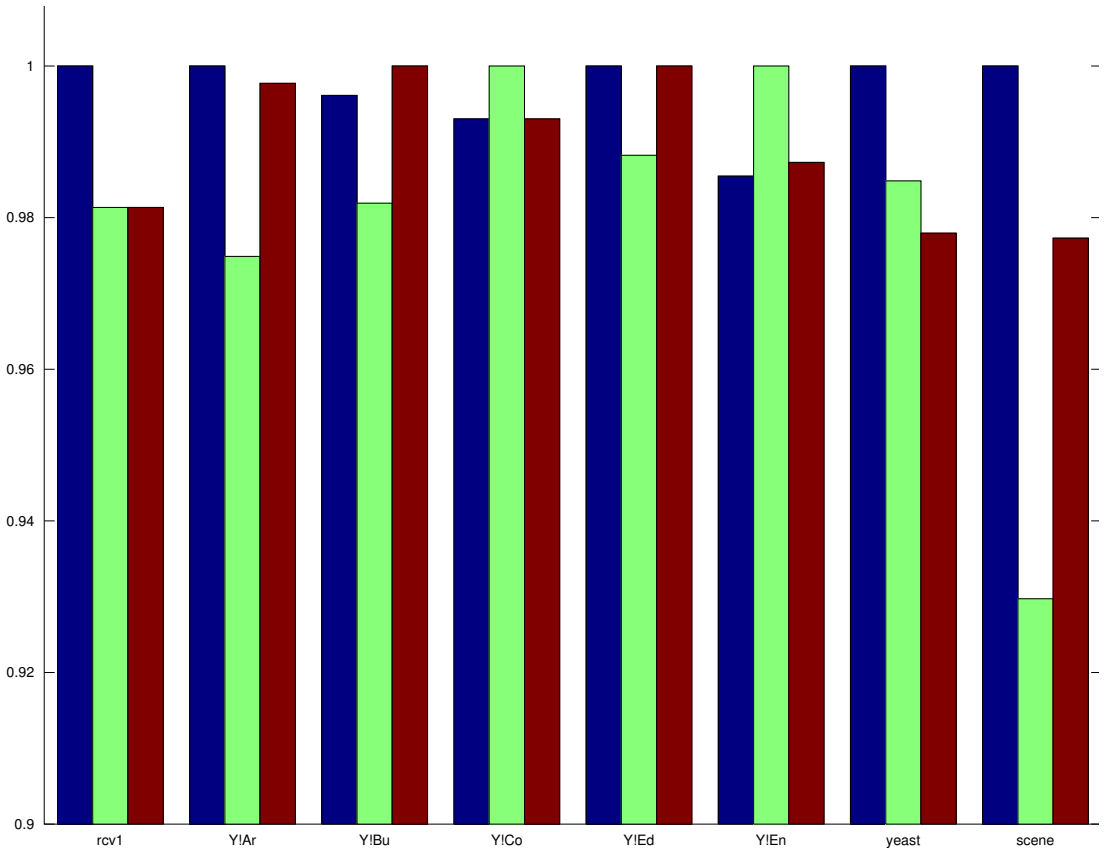
improve MMC by SHLR or HLR?
best criterion **across major/auxiliary combinations**?

SLR+BR, rcv1, Evaluated with F1-score



SHLR > MMR ≈ HLR > BinMin > Random

SLR+BR across Data Sets, Evaluated with F1-score



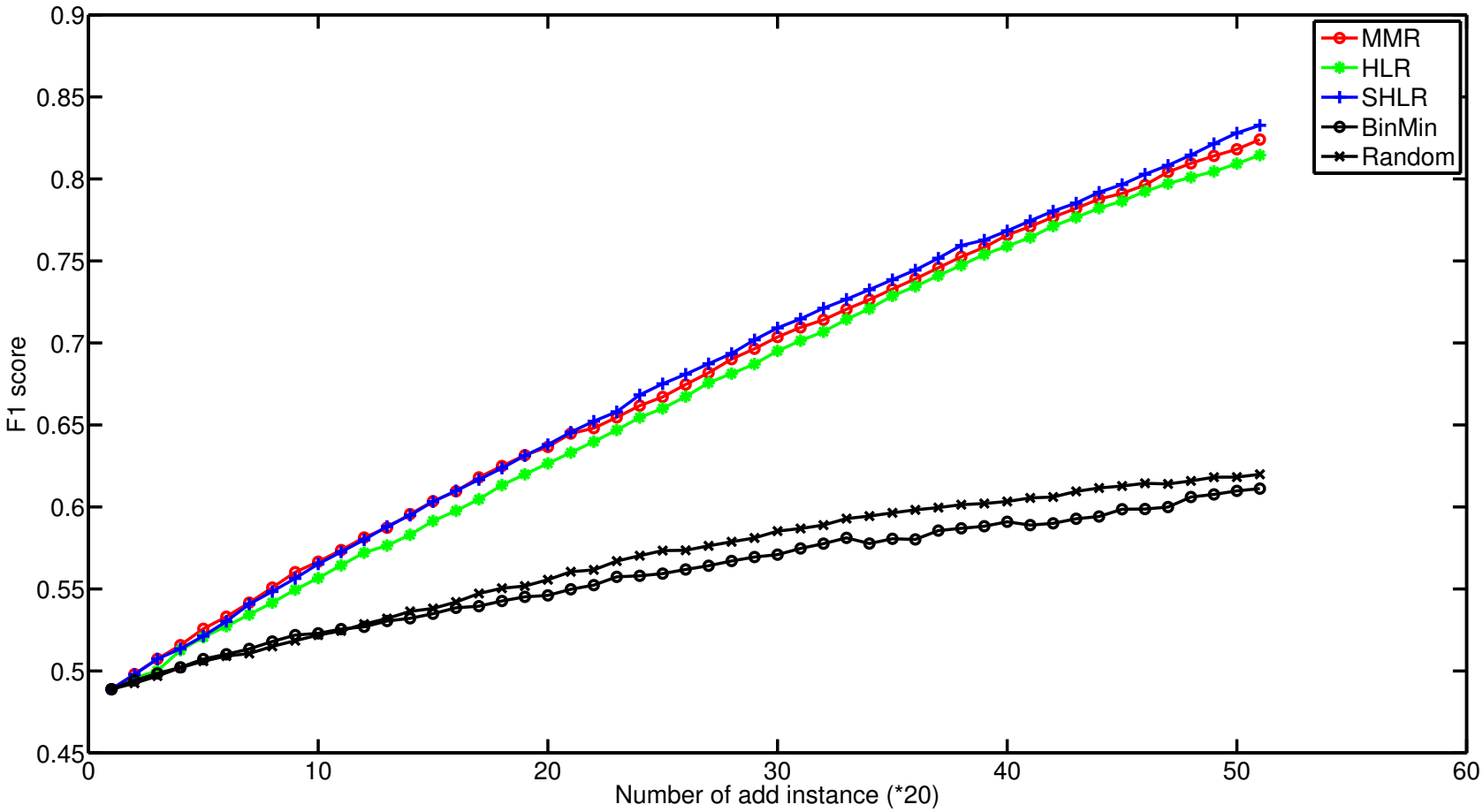
- SHLR best: 5/8 (one tie)
- MMR best: 2/8 (one tie)
- HLR best: 2/8

relative performance to the best across data sets:

SHLR > MMR ≈ HLR

better than MMC? YES!

CC+BR, rcv1, Evaluated with F1-score



SHLR similarly best **when changing major to CC**
—or changing auxiliary to CC
—or changing performance measure to Hamming loss

Conclusion

- general framework for multi-label active learning:
  with auxiliary learner
- simple query criterion:
  via **Hamming loss reduction**, sometimes better
- even better query criterion:
  via **soft Hamming loss reduction**, usually best
- future work:
  major/auxiliary combination, especially choice of auxiliary

Thank you. Questions?