Decision Tree and Instance-Based Learning for Label Ranking

Weiwei Cheng, Jens Hühn and Eyke Hüllermeier

Knowledge Engineering & Bioinformatics Laboratory

Department of Mathematics and Computer Science

Marburg University, Germany

Label Ranking

	label ranking
customer 1	MINI ≻ Toyota ≻ BMW ≻ Volvo
customer 2	BMW ≻ MINI ≻ Toyota
customer 3	Volvo ≻ BMW ≻ Toyota ≻ MINI
customer 4	Toyota ≻ BMW
new customer	???

Given:

- a set of training instances $\{\mathbf{x}_k \mid k = 1 \ldots m\} \subseteq \mathbf{X}$
- a set of labels $\mathcal{L} = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$
- for each training instance $\mathbf{x}_k$: a set of pairwise preferences of the form $\lambda_i \succ_{\mathbf{x}_k} \lambda_j$

Find:

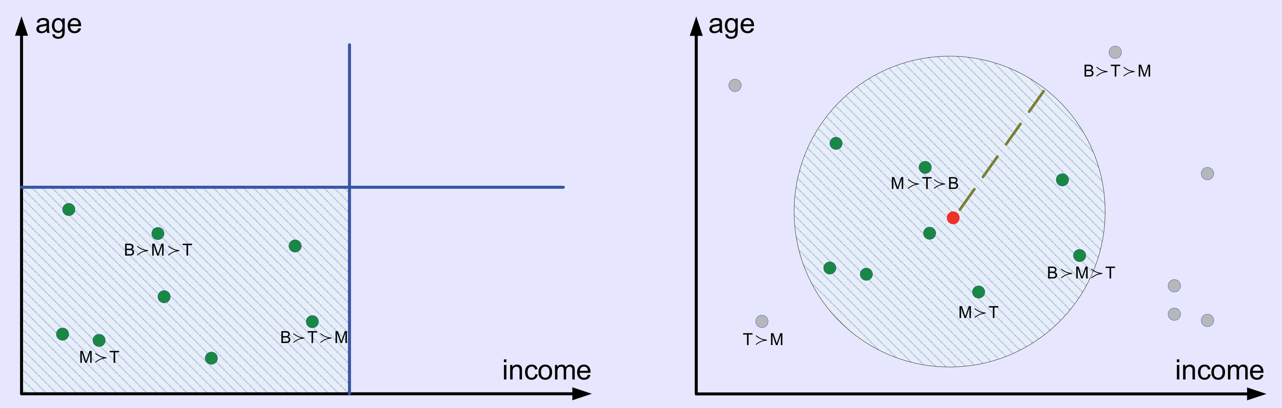
A ranking function ($\mathbf{X} \rightarrow \Omega$ mapping) that maps each $\mathbf{x} \in \mathbf{X}$ to a ranking $\succ_{\mathbf{x}}$ of $\mathcal{L}$ (permutation $\pi_{\mathbf{x}}$) and generalizes well in terms of a loss function on rankings.

Existing Methods

- Ranking by pairwise comparison
  Fürnkranz and Hüllermeier, ECML 2003
- Constraint classification
  Har-Peled, Roth and Zimak, NIPS 2003
- Log linear models for label ranking
  Dekel, Manning and Singer, NIPS 2003

- essentially reduce label ranking to classification
- are efficient but may come with a loss of information
- may have an improper bias and lack flexibility
- or may produce models that are not easily interpretable

Local Learning Approach



- Target function is estimated (on demand) in a local way.
- Core part is to estimate a *locally constant* model.
- Use probabilistic models for rankings, considering nearby preferences as a representative sample.

Mallows Model

$$\mathcal{P}(\sigma \mid \theta, \pi) = \frac{\exp(-\theta d(\pi, \sigma))}{\phi(\theta, \pi)}$$

with *center* $\pi \in \Omega$, *spread* $\theta > 0$, and distance d on Ω.
Maximum likelihood estimation (MLE) based on observed (incomplete) rankings $\boldsymbol{\sigma} = \{\sigma_1, \ldots, \sigma_k\}$:

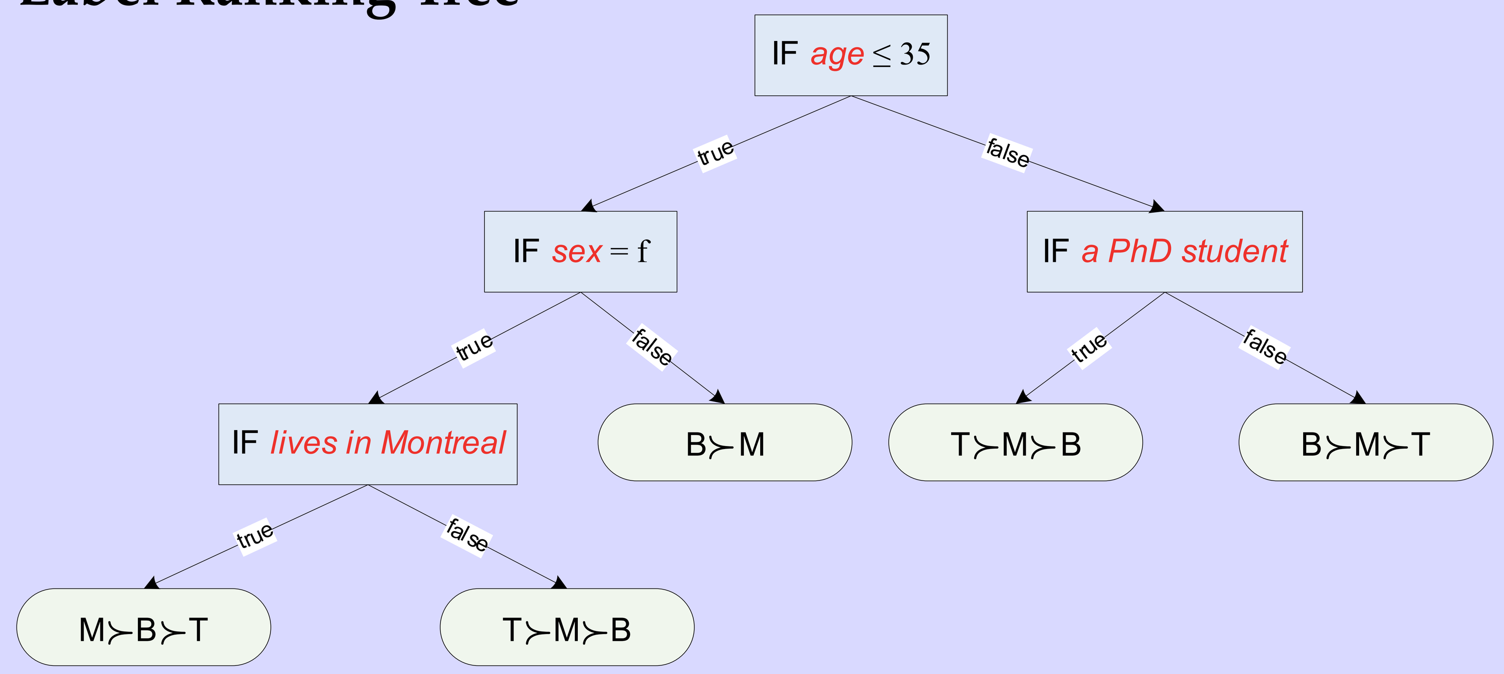
$$\mathcal{P}(\boldsymbol{\sigma} \mid \theta, \pi) = \prod_{i=1}^{k} \mathcal{P}(E(\sigma_i) \mid \theta, \pi) = \prod_{i=1}^{k} \sum_{\sigma \in E(\sigma_i)} \mathcal{P}(\sigma \mid \theta, \pi)$$
$$= \frac{\prod_{i=1}^{k} \sum_{\sigma \in E(\sigma_i)} \exp\left(-\theta d(\sigma, \pi)\right)}{\left(\prod_{j=1}^{n} \frac{1-\exp(-j\theta)}{1-\exp(-\theta)}\right)^k},$$

where $E(\sigma_i)$ denotes the set of consistent extensions of σ_i.

Observation σ	Extensions $E(\sigma)$
$a \succ b$	$a \succ b \succ c$ $a \succ c \succ b$ $c \succ a \succ b$

Approximation of the MLE $(\hat{\pi}, \hat{\theta}) = \arg\max_{\pi,\theta} \mathcal{P}(\boldsymbol{\sigma} \mid \theta, \pi)$ with an EM-like estimation procedure.

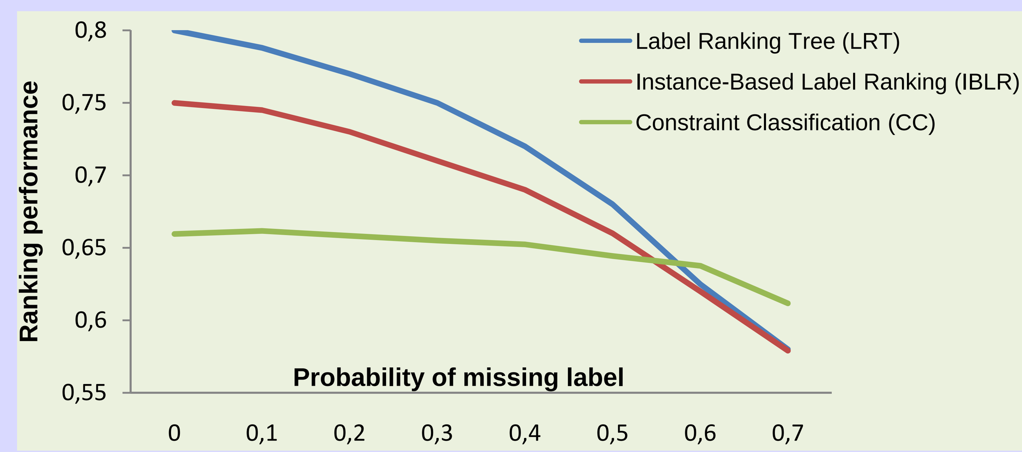
Label Ranking Tree



Major modifications compared with regression trees:

1. Split criterion: seeking a split of T (set of rankings) into T^+ and T^- that maximizes
$$\frac{|T^+| \cdot \theta^+ + |T^-| \cdot \theta^-}{|T|}$$
2. Stopping criterion: tree is pure *OR* number of labels in a node is too small

Main Conclusions from Experiments



- Local learning is more flexible and can exploit more preference information.
- Given enough data, IBLR is significantly better than LRT and CC in terms of predictive accuracy (Kendall's tau).
- The size of LRT is smaller than expected.

ACKNOWLEDGEMENTS: Weiwei Cheng and Jens Hühn were financially supported by the ICML 2009 Student Scholarship Program.

Poster presented at ICML 2009, 26th International Conference on Machine Learning, Montreal, Canada, June 2009