LABEL RANKING METHODS BASED ON THE PLACKETT-LUCE MODEL

Weiwei Cheng, Krzysztof Dembczyński, and Eyke Hüllermeier
Knowledge Engineering & Bioinformatics Lab
Department of Mathematics and Computer Science
University of Marburg, 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 $\{\, \mathbf{x}_i \mid i = 1 \dots N \,\} \subseteq \mathbf{X}$

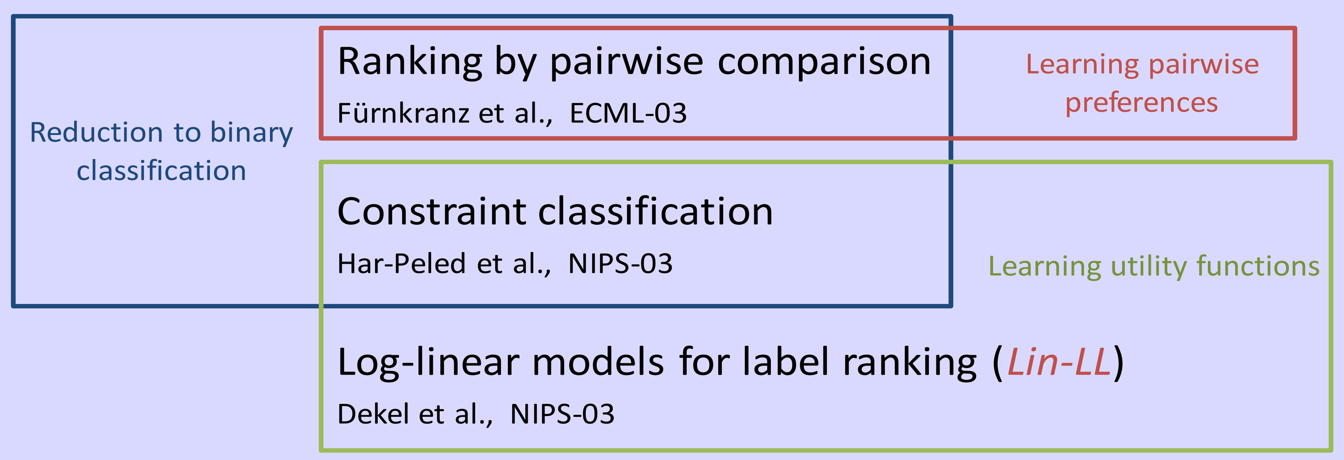
– a set of labels $\mathcal{L} = \{\lambda_1, \lambda_2, \dots, \lambda_M\}$

– for each training instance $\mathbf{x}_i$: a set of pairwise preferences of the form $\lambda_i \succ_{\mathbf{x}_i} \lambda_j$

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



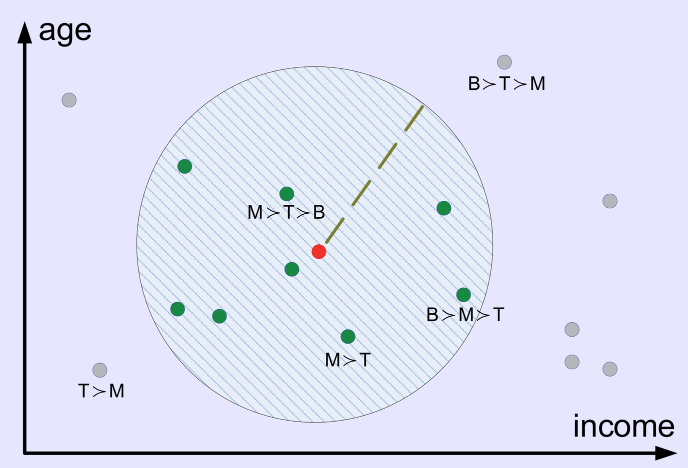
- e.g., Lin-LL minimizes a convex upper bound of the loss

$$\sum_{1 \le i \le j \le M} \begin{cases} 0 & f_{\pi(i)}(\mathbf{x}) < f_{\pi(j)}(\mathbf{x}) \\ 1 & f_{\pi(i)}(\mathbf{x}) \ge f_{\pi(j)}(\mathbf{x}) \end{cases},$$

namely $\log \left[1 + \sum_{1 \le i \le j \le M} \exp\left(f_{\pi(j)}(\mathbf{x}) - f_{\pi(i)}(\mathbf{x})\right) \right]$;

- These methods may have an improper bias and lack flexibility.

Instance-Based Approach



– Target function is estimated (on demand) in a local way;

– Core part is to estimate a locally constant model;

– Uses probabilistic models for rankings, considering nearby preferences as a representative sample.

Plackett-Luce Model

$$\mathcal{P}(\Pi = \pi; \mathbf{v}) = \prod_{i=1}^{M} \frac{v_{\pi(i)}}{v_{\pi(i)} + v_{\pi(i+1)} + \dots + v_{\pi(M)}}$$

– Positive $v_1, \dots, v_M$, where v_i corresponds to i-th label's score, ability, skill, etc.;

– First determines the 1st rank, then the 2nd rank, and so on (i.e., a *multistage model*);

– Appealing for incomplete rankings. The probability of an incomplete ranking with $k < M$ labels observed: $\mathcal{P}(\Pi = \pi; \mathbf{v}) = \prod_{i=1}^{k} v_{\pi(i)} \,/\, \left(v_{\pi(i)} + v_{\pi(i+1)} + \dots + v_{\pi(k)}\right)$.

- The probability to observe rankings $\boldsymbol{\pi} = \{\pi_1 \dots \pi_K\}$ in the neighborhood: $\mathcal{P}(\boldsymbol{\pi}; \mathbf{v}) = \prod_{i=1}^{K} \prod_{m=1}^{M_i} v_{\pi_i(m)} / \left(\sum_{j=m}^{M_i} v_{\pi_i(j)}\right)$;
- Corresponding MLE can be done efficiently, e.g., through MM (minorization and maximization) algorithm. See *MM Algorithm for Generalized Bradley-Terry Models*, Hunter, The Annals of Statistics, 2004.

Generalized Linear Model

– Modeling the parameter v_i as a linear function of the attributes describing the instance: $v_i = \exp\left(\sum_{d=1}^{D} \alpha_d^{(i)} \cdot x_d\right)$, $1 \le i \le M, 1 \le d \le D$;

– Given the training data $\mathcal{T} = \left\{\left(\mathbf{x}^{(n)}, \pi^{(n)}\right)\right\}_{n=1}^{N}$ with $\mathbf{x}^{(n)} = \left(x_1^{(n)} \dots x_D^{(n)}\right)$, the log-likelihood function is

$$L = \sum_{n=1}^{N} \left[\sum_{i=1}^{M_n} \log\left(v(\pi^{(n)}(i), n)\right) - \log \sum_{j=m}^{M_n} v(\pi^{(n)}(j), n) \right],$$

where M_n is the number of labels in the ranking $\pi^{(n)}$ and $v(i,n) = \exp\left(\sum_{d=1}^{D} \alpha_d^{(i)} \cdot x_d^{(n)}\right)$;

– L is convex with respect to $\alpha_d^{(i)}$.

Experiments and Conclusions

	complete ranking				30% missing labels				60% missing labels			
	IB-PL	IB-Mal	Lin-PL	Lin-LL	IB-PL	IB-Mal	Lin-PL	Lin-LL	IB-PL	IB-Mal	Lin-PL	Lin-LL
authorship	.936(1)	.936(2)	.930(3)	.657(4)	.927(1)	.913(2)	.899(3)	.656(4)	.886(1)	.849(2)	.846(3)	.650(4)
bodyfat	.230(3)	.229(4)	.272(1)	.266(2)	.204(3)	.198(4)	.266(1)	.251(2)	.151(4)	.160(3)	.222(2)	.241(1)
calhousing	.326(2)	.344(1)	.220(4)	.223(3)	.303(2)	.310(1)	.229(3)	.223(4)	.259(2)	.263(1)	.229(3)	.221(4)
cpu-small	.495(2)	.496(1)	.426(3)	.419(4)	.477(1)	.473(2)	.418(4)	.419(3)	.437(1)	.428(2)	.412(4)	.418(3)
elevators	.721(2)	.727(1)	.712(3)	.701(4)	.702(2)	.683(4)	.706(1)	.699(3)	.633(3)	.596(4)	.704(1)	.696(2)
fried	.894(4)	.900(3)	.996(1)	.989(2)	.861(3)	.850(4)	.993(1)	.989(2)	.797(3)	.777(4)	.990(1)	.987(2)
glass	.841(2)	.842(1)	.825(3)	.818(4)	.809(3)	.776(4)	.825(1)	.817(2)	.675(3)	.611(4)	.807(2)	.808(1)
housing	.711(2)	.736(1)	.659(3)	.626(4)	.654(3)	.669(1)	.658(2)	.625(4)	.492(4)	.543(3)	.636(1)	.614(2)
iris	.960(1)	.925(2)	.832(3)	.818(4)	.926(1)	.867(2)	.823(3)	.804(4)	.868(1)	.799(2)	.778(3)	.768(4)
pendigits	.939(2)	.941(1)	.909(3)	.814(4)	.918(1)	.902(3)	.909(2)	.802(4)	.794(2)	.781(4)	.907(1)	.787(3)
segment	.950(1)	.802(4)	.902(2)	.810(3)	.874(2)	.735(4)	.895(1)	.806(3)	.674(3)	.612(4)	.888(1)	.801(2)
stock	.922(2)	.925(1)	.710(3)	.696(4)	.877(1)	.855(2)	.701(3)	.691(4)	.740(1)	.724(2)	.687(4)	.689(3)
vehicle	.859(1)	.855(2)	.838(3)	.770(4)	.838(1)	.822(2)	.817(3)	.769(4)	.765(2)	.736(4)	.804(1)	.764(3)
vowel	.851(2)	.882(1)	.586(4)	.601(3)	.785(2)	.810(1)	.581(4)	.598(3)	.588(3)	.638(1)	.575(4)	.591(2)
wine	.947(2)	.944(3)	.954(1)	.942(4)	.926(4)	.930(3)	.931(2)	.941(1)	.907(2)	.893(4)	.915(1)	.894(3)
wisconsin	.479(4)	.501(3)	.635(1)	.542(2)	.453(4)	.464(3)	.615(1)	.533(2)	.381(4)	.399(3)	.585(1)	.518(2)
Avg. Rank	2.06	1.94	2.56	3.44	2.13	2.63	2.19	3.06	2.44	2.94	2.06	2.56

– Instance-based methods are more flexible, while linear methods are more robust;

– Probabilistic modeling of the data generating process leads to a theoretically sound method and has further advantages compared to direct loss minimization.

ACKNOWLEDGEMENT: Weiwei Cheng was financially supported by the ICML 2010 Student Travel Scholarship Program.
Poster presented at ICML 2010, 27th International Conference on Machine Learning, Haifa, Israel, June 2010