Label Ranking Methods based on the Plackett-Luce Model

Weiwei Cheng, Krzysztof Dembczynski, Eyke Hüllermeier

Knowledge Engineering & Bioinformatics Lab
Department of Mathematics and Computer Science
University of Marburg, Germany

Philipps Universität Marburg

Label Ranking – An Example

Learning customers' preferences on cars:

	label ranking
customer 1	MINI $\succ$ Toyota $\succ$ BMW
customer 2	BMW $\succ$ MINI $\succ$ Toyota
customer 3	BMW $\succ$ Toyota $\succ$ MINI
customer 4	Toyota $\succ$ MINI $\succ$ BMW
new customer	???

where the customers could be described by feature vectors, e.g., (gender, age, place of birth, has child, …)

Label Ranking – An Example

Learning customers' preferences on cars:

	MINI	Toyota	BMW
customer 1	1	2	3
customer 2	2	3	1
customer 3	3	2	1
customer 4	2	1	3
new customer	?	?	?

$\pi(i)$ = position of the *i*-th label in the ranking

1: MINI 2: Toyota 3: BMW

Existing Approaches

Reduction to binary classification

Ranking by pairwise comparison
Fürnkranz et al., ECML-03

Learning pairwise preferences

Constraint classification
Har-Peled et al., NIPS-03

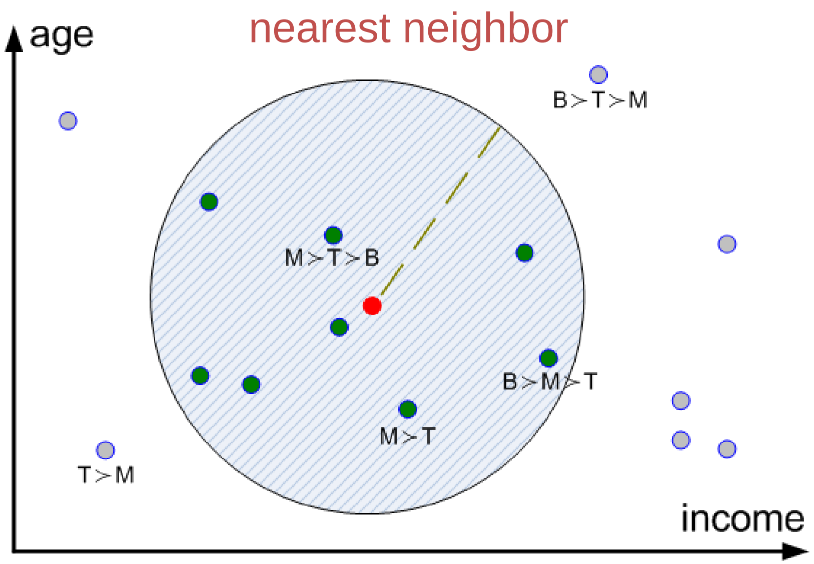
Learning utility functions

Log-linear models for label ranking (*Lin-LL*)
Dekel et al., NIPS-03

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

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

Instance-Based Approaches



- Target function $\mathcal{X} \rightarrow \Omega$ is estimated (on demand) in a local way.
- Distribution of rankings is (approx.) constant in a local region.
- Core part is to estimate the locally constant model.

Ranking with Mallows Model

Cheng et al. ICML-09

Mallows model (Mallows, Biometrika, 1957)

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

with

center ranking $\pi \in \Omega$

spread parameter $\theta > 0$

and $d(\cdot)$ is a metric on permutations

Computational issues arise when the training data contains incomplete rankings.

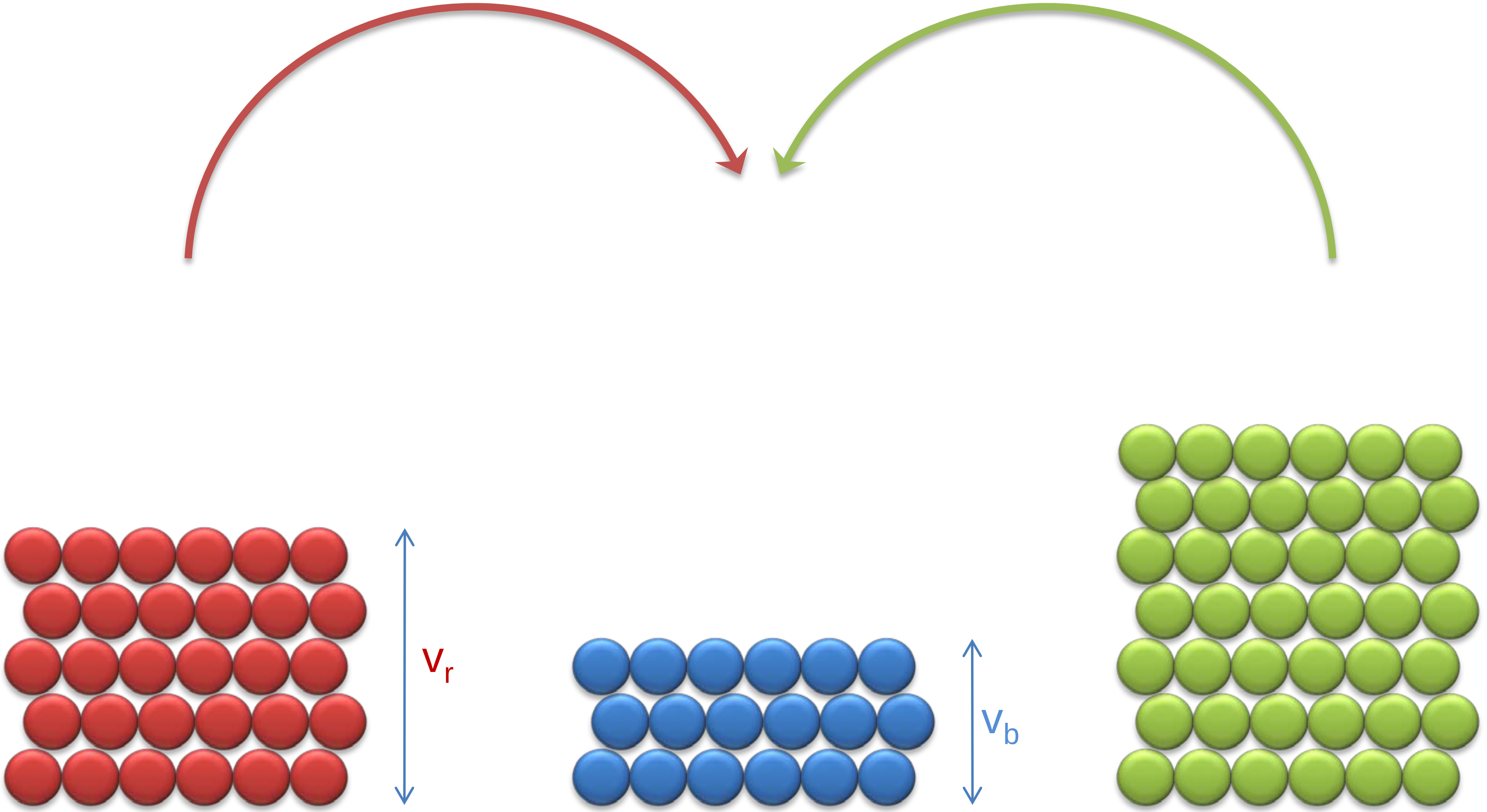
$$\mathcal{P}(\pi \,|\, \theta, \pi_0) = \sum_{\pi^* \in E(\pi)} \mathcal{P}(\pi^* \,|\, \theta, \pi_0)$$

Multistage model

- First determine the 1st rank, then the 2nd rank, etc.
- Positive $v_1, \ldots, v_M$, where v_i corresponds to *i*-th label's score, ability, skill, etc.
- Plackett-Luce

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

PL: Vase Interpretation



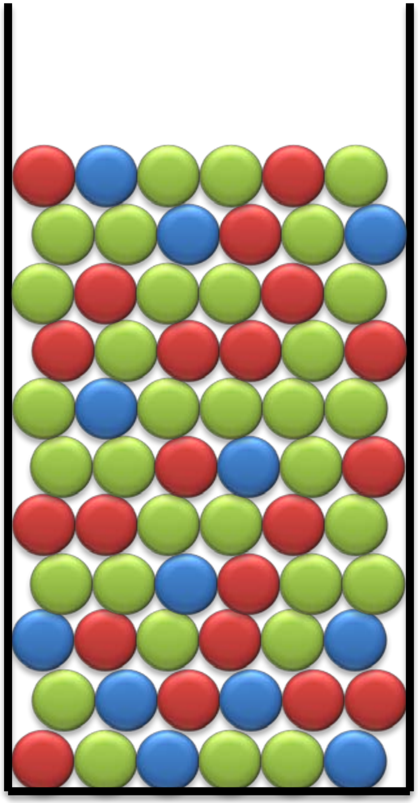
PL: Vase Interpretation

Probability:

$$\frac{v_r}{v_r + v_g + v_b}$$

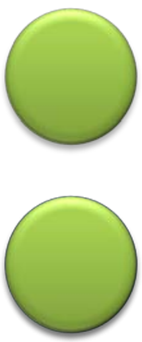
$$\times \frac{v_g}{v_g + v_b}$$

$$\times \frac{v_b}{v_b}$$



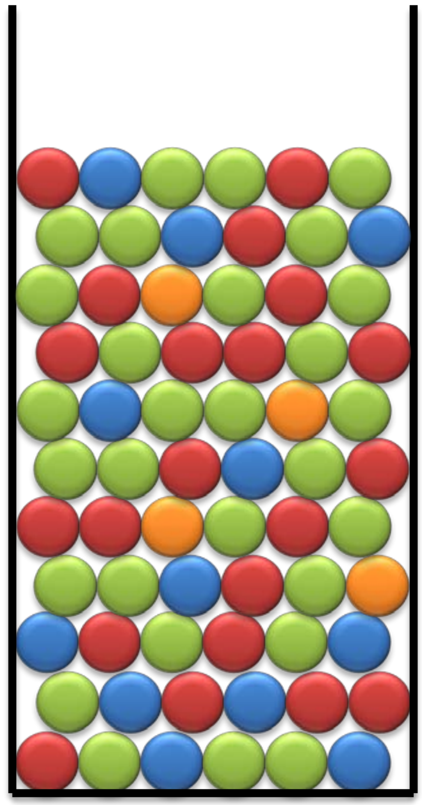
PL: Vase Interpretation

Top K

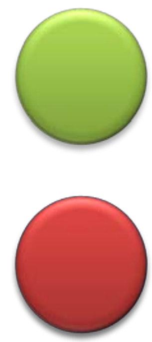


$$\frac{v_g}{v_r + v_g + v_b + v_o}$$

$$\times \frac{v_b}{v_r + v_b + v_o}$$



Incomplete ranking



$$\frac{v_g}{v_g + v_b}$$

$$\times \frac{v_b}{v_b}$$

(Bradley-Terry model for case of pairs)

Multistage Model

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

- First determine the 1st rank, then the 2nd rank, etc.

- Plackett-Luce

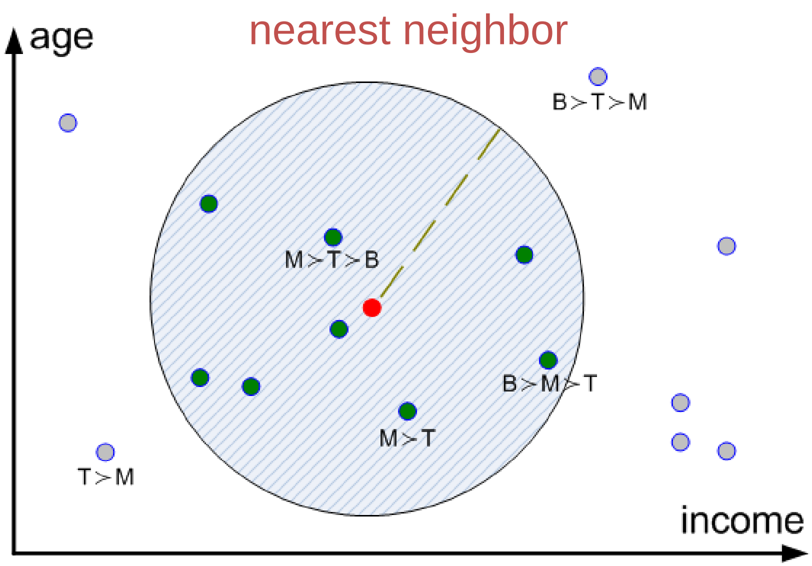
$$\mathcal{P}(\Pi = \pi; \mathbf{v}) = \prod_{i=1}^{M} \frac{v_{\pi(i)}}{v_{\pi(i)} + v_{\pi(i+1)} + \cdots + v_{\pi(M)}}$$

- For the incomplete ranking

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

$k < M$ is the number of labels observed.

Instance-Based Label Ranking



The probability to observe the rankings $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_K\}$ in the neighborhood:

$$\mathcal{P}(\boldsymbol{\pi}; \mathbf{v}) = \prod_{i=1}^{K} \prod_{m=1}^{M_i} \frac{v_{\pi_i(m)}}{\sum_{j=m}^{M_i} v_{\pi_i(j)}}$$

Corresponding MLE can be efficiently done through, e.g., MM (minorization and maximization) algorithm, see Hunter 2004.

Lazy approach

Generalized Linear Models

Can we take advantage of the global approaches?

Low variance, highly stable w.r.t. runtime, performance, etc.

- Estimating a *global* model
- Modeling the parameter v_i as a linear function of the attributes describing the instance.

$$v_i = \exp\left(\sum_{d=1}^{D} \alpha_d^{(i)} \cdot x_d\right), \ (1 \leq i \leq M, 1 \leq d \leq D)$$

Generalized Linear Models

Given training data $\mathcal{T} = \left\{ \left(\mathbf{x}^{(n)}, \pi^{(n)} \right) \right\}_{n=1}^{N}$ with $\mathbf{x}^{(n)} = \left(x_1^{(n)}, \ldots, x_D^{(n)} \right)$,
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

$$v(i, n) = \exp \left(\sum_{d=1}^{D} \alpha_d^{(i)} \cdot x_d^{(n)} \right).$$

This log-likelihood function is convex!

Experimental Results

	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

IB-PL: instance-based with PL
IB-Mal: instance-based with Mallows
Lin-PL: linear model with PL
Lin-LL: log linear approach

Performance measured in *Kendall's tau*

(#concordant pairs - #discordant pairs) / #all pairs

Experimental Results

	IB-PL	IB-Mal	Lin-PL	Lin-LL
IB-PL	—	6/11/11	12/8/7	13/11/9
IB-Mal	10/5/5	—	11/8/7	12/9/7
Lin-PL	4/8/9	5/8/9	—	14/13/11
Lin-LL	3/5/7	4/7/9	2/4/5	—

win/win/win statistics for complete rankings, 30% and 60% missing labels

IB-PL: instance-based with PL

IB-Mal: instance-based with Mallows

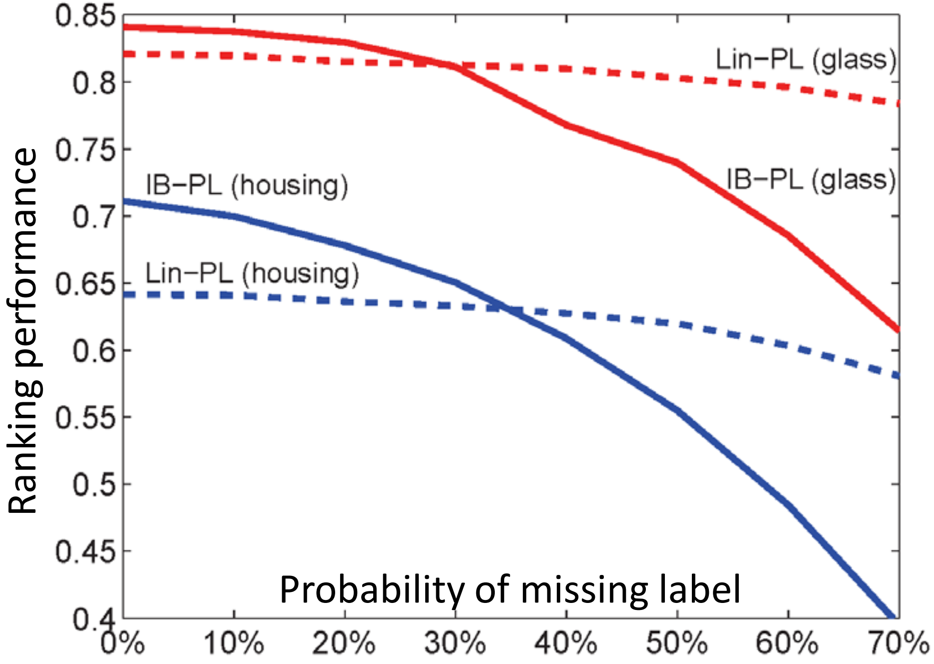
Lin-PL: linear model with PL

Lin-LL: log linear approach

Performance measured in *Kendall's tau*

(#concordant pairs - #discordant pairs) / #all pairs

Typical “learning curves”



Main observation:

Local methods are more *flexible* and can exploit more preference information; global approaches are more *robust*.

Summary

- Label ranking with *Plackett-Luce* Model
  - Instance-based approach
  - Generalized linear approach
- Particularly appealing for training with incomplete ranking

- Probabilistic modeling of the data generating process
- Some advantages compared to direct loss minimization

- Combining local and global methods, estimating a linear model in a local way

Knowledge Engineering & Bioinformatics (KEBI)
Mathematics and Computer Science
University of Marburg

http://www.uni-marburg.de/fb12/kebi