A Simple Instance-Based Approach to Multilabel Classification Using the Mallows Model

Weiwei Cheng & Eyke Hüllermeier

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

Label Ranking (an example)

Learning visitor's preferences on hotels

	label ranking
visitor 1	Golf ≻ Park ≻ Krim
visitor 2	Krim ≻ Golf ≻ Park
visitor 3	Krim ≻ Park ≻ Golf
visitor 4	Park ≻ Golf ≻ Krim
new visitor	???

where the visitor could be described by feature vectors,
e.g., (*gender*, *age*, *place of birth*, *is a professor*, ...)

Label Ranking (an example)

Learning visitor's preferences on hotels

	Golf	Park	Krim
visitor 1	1	2	3
visitor 2	2	3	1
visitor 3	3	2	1
visitor 4	2	1	3
new visitor	?	?	?

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

1: Golf 2: Park 3: Krim

Label Ranking (more formally)

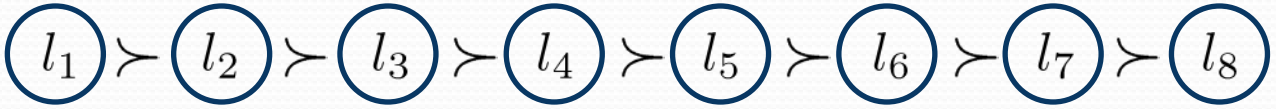
Given:

- a set of training instances $\{\mathbf{x}_k \mid k = 1 \ldots m\} \subseteq \mathbf{X}$
- a set of labels $\mathcal{L} = \{l_i \mid i = 1 \ldots n\}$
- for each training instance $\mathbf{x}_k$: a set of *pairwise preferences* of the form $l_i \succ_{\mathbf{x}_k} l_j$ (for some of the labels)

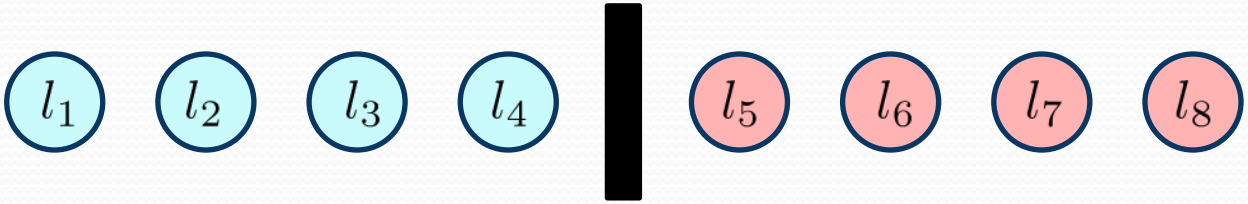
Find:

- A ranking function ($\mathcal{X} \rightarrow \Omega$ 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 (e.g., *Kendall's tau*)

Label ranking

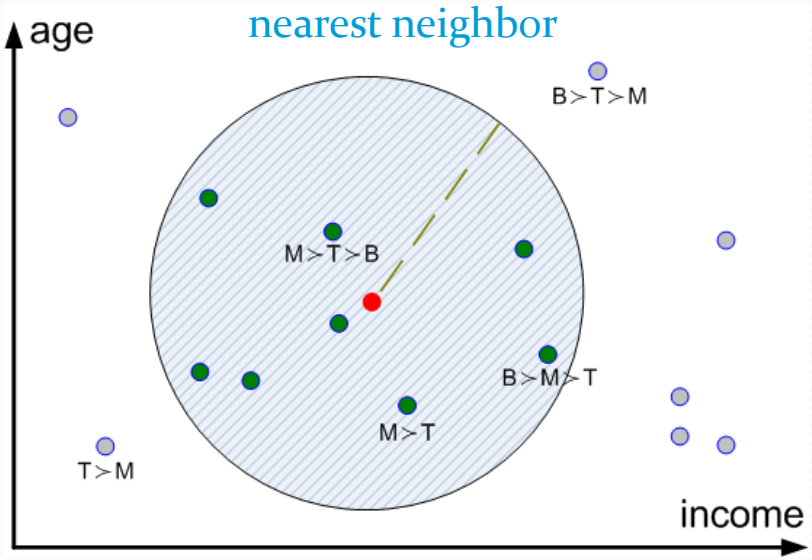


Multilabel classification



Calibrated label ranking

Instance-based Label Ranking



- 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.

Probabilistic Model for Ranking

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

Inference

$$L_{x_1} = \{l_1, l_2\} \succ l_0 \succ \{l_3\}$$

$$L_{x_3} = \{l_2\} \succ l_0 \succ \{l_1, l_3\}$$

$$L_{x_2} = \{l_1\} \succ l_0 \succ \{l_2, l_3\}$$

$$\mathcal{P}(L_{x_i} \,|\, \theta, \pi) = \sum_{\sigma \in E(L_{x_i})} \mathcal{P}(\sigma \,|\, \theta, \pi)$$

An example:

Observation: $\{l_1, l_2\} \succ l_0 \succ \{l_3\}$

Extensions:

$$l_1 \succ l_2 \succ l_0 \succ l_3$$

$$l_2 \succ l_1 \succ l_0 \succ l_3$$

Inference (Cont.)

For observations $\mathbf{L} = \{L_{x_1} \dots L_{x_k}\}$:

$$\begin{aligned}
\mathcal{P}(\mathbf{L} \mid \theta, \pi) &= \prod_{i=1}^{k} \mathcal{P}(E(L_{x_i}) | \theta, \pi) \\
&= \prod_{i=1}^{k} \sum_{\sigma \in E(L_{x_i})} \mathcal{P}(\sigma | \theta, \pi) \\
&= \frac{\prod_{i=1}^{k} \sum_{\sigma \in E(L_{x_i})} \exp\left(-\theta d(\sigma, \pi)\right)}{\left(\prod_{j=1}^{n} \frac{1-\exp(-j\theta)}{1-\exp(-\theta)}\right)^{k}}.
\end{aligned}$$

Maximum Likelihood Estimation (MLE) becomes a difficult, non-convex optimization problem!

Special Structure of Observations

In the multilabel classification context, the observation is a special type of *partial ranking*. It contains:

- two tie groups (i.e. relevant & irrelevant label set)
- all labels (i.e. no label is missing)

By exploiting this special structure , MLE can be made much more efficiently!

First Theorem

Theorem 1: For each label $\lambda_i \in \mathcal{L}$, let $f(\lambda_i)$ denote the frequency of occurrence of this label in the neighborhood of $\boldsymbol{x}$, i.e., $f(\lambda_i) = \#\{j \,|\, \lambda_i \in L_{\boldsymbol{x}_j}\}/k$. Moreover, let $f(\lambda_0) = 1/2$ by definition. Then, a ranking $\pi \in \Omega$ is a MLE in (3) iff it guarantees that $f(\lambda_i) > f(\lambda_j)$ implies $\pi(i) < \pi(j)$.

Sorting the labels according to their frequency of occurrence in the neighborhood

Problem: What about *ties*?
Solution: Replacing MLE with *Bayes estimation*.

Second Theorem

Theorem 2: Let $g(\lambda_i)$ denote the frequency of occurrence of the label λ_i in the complete training set. There exists a prior distribution $\mathbf{P}$ on Ω such that, for large enough k, a ranking $\pi \in \Omega$ is a maximum posterior probability (MAP) estimation iff it guarantees the following: If $f(\lambda_i) > f(\lambda_j)$ or $f(\lambda_i) = f(\lambda_j)$ and $g(\lambda_i) > g(\lambda_j)$, then $\pi(i) < \pi(j)$.

A simple prediction procedure:

1. Sort labels with their frequency in the neighborhood
2. Break ties with frequency outside

Experimental Setting

dataset	domain	#inst.	#attr.	#labels	card.
emotions	music	593	72	6	1,87
image	vision	2000	135	5	1,24
genbase	biology	662	1186(n)	27	1,25
mediamill	multimedia	5000	120	101	4,27
reuters	text	7119	243	7	1,24
scene	vision	2407	294	6	1,07
yeast	biology	2417	103	14	4,24

Tested methods:

- MLKNN
- binary relevance learning (BR) with C4.5
- Our method: Mallows

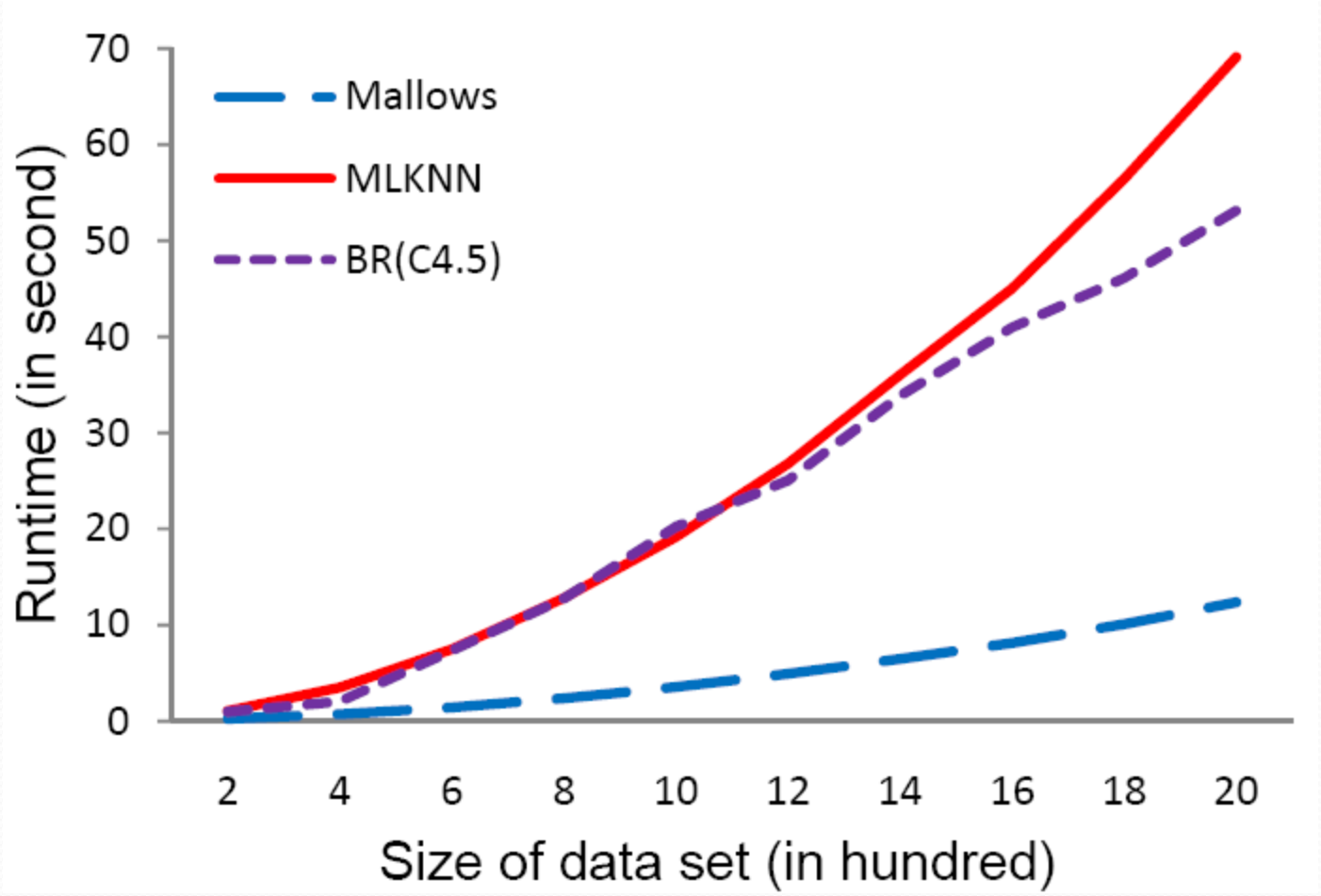
Evaluation metrics

- *Hamming loss* and *rank loss*

Experimental Results

	Hamming loss			rank loss		
dataset	BR	MLKNN	Mallows	BR	MLKNN	Mallows
emotions	0.253	0.261	0.197	0.352	0.262	0.163
image	0.001	0.005	0.003	0.006	0.006	0.006
genbase	0.243	0.193	0.192	0.398	0.214	0.208
mediamill	0.032	0.027	0.027	0.189	0.037	0.036
reuters	0.057	0.073	0.085	0.089	0.068	0.087
scene	0.131	0.087	0.094	0.300	0.077	0.088
yeast	0.249	0.194	0.197	0.360	0.168	0.165

Experimental Results (Cont.)



Our Contributions

- A new instance-based multilabel classifier with state-of-the-art predictive accuracy;
- It is computationally very efficient;
- with very simple prediction procedure, justified by an underlying probabilistic ranking model.

The End

Thanks!

Google “kebi germany” for more info.