A New Instance-Based Label Ranking Approach 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 customers’ preferences on cars:

<table>
<thead>
<tr>
<th>label ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>customer 1</td>
</tr>
<tr>
<td>customer 2</td>
</tr>
<tr>
<td>customer 3</td>
</tr>
<tr>
<td>customer 4</td>
</tr>
<tr>
<td>new customer</td>
</tr>
</tbody>
</table>

where the customers are typically described by feature vectors, e.g., (gender, age, place of birth, has child, ...
### Label Ranking (an example)

Learning customers’ preferences on cars:

<table>
<thead>
<tr>
<th></th>
<th>MINI</th>
<th>Toyota</th>
<th>BMW</th>
</tr>
</thead>
<tbody>
<tr>
<td>customer 1</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>customer 2</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>customer 3</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>customer 4</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>new customer</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

\[
\pi(i) = \text{position of the } i\text{-th label in the ranking}
\]

1: MINI \quad 2: Toyota \quad 3: BMW
Label Ranking (more formally)

Given:

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

Find:

- a ranking function \( (X \to \Omega \text{ mapping}) \) that maps each \( x \in X \) to a ranking \( \succ_x \) of \( L \) (permutation \( \pi_x \))
Model-Based Approaches

... essentially reduce label ranking to classification:

- **Ranking by pairwise comparison (RPC)**
  Fürnkranz and Hüllermeier, ECML-03

- **Constraint classification (CC)**
  Har-Peled, Roth and Zimak, NIPS-03

- **Log linear models for label ranking (LL)**
  Dekel, Manning and Singer, NIPS-03
Instance-Based Approach (this work)

- Lazy learning: Instead of eagerly inducing a model from the data, simply store the observations.

- Target functions are estimated on demand in a local way, no need to define the \( x \rightarrow \Omega \) mapping explicitly.

- Core part is the aggregation of preference (order) information from neighbored examples.
Related Work

*Case-based Label Ranking* (Brinker and Hüllermeier, ECML-06)

Aggregation of *complete* rankings is done by
- median ranking
- Borda count

Our aggregation method is based on a probabilistic model and can handle both *complete* and *incomplete* rankings.
Instance-Based Prediction

Basic assumption: Distribution of output is (approximately) constant in the neighborhood of the query; consider outputs of neighbors as an i.i.d. sample.

Conventional classification:
- discrete distribution on class labels
- estimate probabilities by relative class frequencies
- class prediction by majority vote
Probabilistic Model for Ranking

Mallows model (Mallows, Biometrika, 1957)

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

with

center ranking \( \pi \in \Omega \)

spread parameter \( \theta > 0 \)

and \( d(\cdot) \) is a right invariant metric on permutations

\[ \forall \pi, \sigma, \nu \in \Omega, \ d(\pi, \sigma) = d(\pi \nu, \sigma \nu). \]
Inference (full rankings)

We have observed $\sigma = \{\sigma_1, \ldots, \sigma_k\}$ from the neighbors.

$$P(\sigma|\theta, \pi) = \prod_{i=1}^{k} P(\sigma_i|\theta, \pi)$$

$$= \prod_{i=1}^{k} \frac{\exp(-\theta d(\sigma_i, \pi))}{\phi(\theta)}$$

$$= \exp\left(-\theta (d(\sigma_1, \pi) + \ldots + d(\sigma_k, \pi))\right)$$

$$= \frac{\exp\left(-\theta \sum_{i=1}^{k} d(\sigma_i, \pi)\right)}{\left(\Pi_{j=1}^{n} \frac{1-\exp(-j\theta)}{1-\exp(-\theta)}\right)^k}.$$
**Inference** (incomplete ranking)

“marginal” distribution \( \mathcal{P}(E(\sigma_i)) = \sum_{\sigma \in E(\sigma_i)} \mathcal{P}(\sigma | \theta, \pi) \)

where \( E(\sigma_i) \) denotes all consistent extensions of \( \sigma_i \).

Example for label set \( \{a, b, c\} \):

<table>
<thead>
<tr>
<th>Observation ( \sigma )</th>
<th>Extensions ( E(\sigma) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a \succ b )</td>
<td>( a \succ b \succ c )</td>
</tr>
<tr>
<td></td>
<td>( a \succ c \succ b )</td>
</tr>
<tr>
<td></td>
<td>( c \succ a \succ b )</td>
</tr>
</tbody>
</table>
Inference (incomplete ranking) cont.

The corresponding likelihood:

\[ \mathcal{P}(\sigma | \theta, \pi) = \prod_{i=1}^{k} \mathcal{P}(E(\sigma_i) | \theta, \pi) \]

\[ = \prod_{i=1}^{k} \sum_{\sigma \in E(\sigma_i)} \mathcal{P}(\sigma | \theta, \pi) \]

\[ = \prod_{i=1}^{k} \sum_{\sigma \in E(\sigma_i)} \exp \left( -\theta d(\sigma, \pi) \right) \]

\[ = \frac{\prod_{i=1}^{k} \sum_{\sigma \in E(\sigma_i)} \exp \left( -\theta d(\sigma, \pi) \right)}{\left( \prod_{j=1}^{n} \frac{1-\exp(-j\theta)}{1-\exp(-\theta)} \right)^k}. \]

ML estimation \( (\hat{\pi}, \hat{\theta}) = \arg \max_{\pi, \theta} \mathcal{P}(\sigma | \theta, \pi) \) becomes more difficult.
Inference

Not only the estimated ranking $\hat{\pi}$ is of interest ...

... but also the spread parameter $\hat{\theta}$, which is a measure of precision and, therefore, reflects the confidence/reliability of the prediction (just like the variance of an estimated mean).

The bigger $\hat{\theta}$, the more peaked the distribution around the center ranking.
Experimental Setting

Data sets

<table>
<thead>
<tr>
<th>Name</th>
<th>#instances</th>
<th>#features</th>
<th>#labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris(^1)</td>
<td>150</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Wine(^1)</td>
<td>178</td>
<td>13</td>
<td>3</td>
</tr>
<tr>
<td>Glass(^1)</td>
<td>214</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>Vehicle(^1)</td>
<td>846</td>
<td>18</td>
<td>4</td>
</tr>
<tr>
<td>Dtt(^2)</td>
<td>2465</td>
<td>24</td>
<td>4</td>
</tr>
<tr>
<td>Cold(^2)</td>
<td>2465</td>
<td>24</td>
<td>4</td>
</tr>
</tbody>
</table>

\(^1\) UCI data sets.

\(^2\) Phylogenetic profiles and DNA microarray expression data.
Accuracy (Kendall tau)

A typical run:

Main observation: Our approach is quite competitive with the state-of-the-art model based approaches.
Accuracy-Rejection Curve

\( \theta \) as a measure of the reliability of a prediction

glass

Main observation: Decreasing curve confirms that \( \theta \) is a reasonable measure of confidence.
Take-away Messages

- An instance-based label ranking approach using a probabilistic model.
- Suitable for complete and incomplete rankings.
- Comes with a natural measure of the reliability of a prediction.
  - More efficient inference for the incomplete case.
  - Generalization: distance-weighted prediction.
  - Dealing with variants of the label ranking problem, such as calibrated label ranking and multi-label classification.
Thanks!
Median rank

<table>
<thead>
<tr>
<th></th>
<th>MINI</th>
<th>Toyota</th>
<th>BMW</th>
</tr>
</thead>
<tbody>
<tr>
<td>ranking 1</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>ranking 2</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>ranking 3</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>median rank</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

- tends to optimize Spearman footrule  \( \sum_{i=1}^{n} |\pi(i) - \hat{\pi}(i)| \)
Borda count

<table>
<thead>
<tr>
<th>ranking 1</th>
<th>MINI $\succ$ Toyota $\succ$ BMW</th>
</tr>
</thead>
<tbody>
<tr>
<td>ranking 2</td>
<td>BMW $\succ$ MINI $\succ$ Toyota</td>
</tr>
<tr>
<td>ranking 3</td>
<td>BMW $\succ$ Toyota $\succ$ MINI</td>
</tr>
<tr>
<td>Borda count</td>
<td>BMW: 4 MINI: 3 Toyota: 2</td>
</tr>
</tbody>
</table>

- tends to optimize Spearman rank correlation $\sum_{i=1}^{n} (\pi(i) - \hat{\pi}(i))^2$