Ranking Median Regression: Learning to Order through Local Consensus

Anna Korba* Stephan Clémençon* Eric Sibony†

* Telecom ParisTech, † Shift Technology

Algorithmic Learning Theory
April 8 2018

Ranking Regression

Consider:

- A set of n items: $[\![n]\!] = \{1, \ldots, n\}$
- A individual expresses her preferences as (full) ranking, i.e a strict order $\succ$ over n :

$$a_1 \succ a_2 \succ \cdots \succ a_n$$

- Also seen as the permutation σ that maps an item to its rank:

$$a_1 \succ \cdots \succ a_n \quad \Leftrightarrow \quad \sigma \in \mathfrak{S}_n \text{ such that } \sigma(a_i) = i$$

$\mathfrak{S}_n$: set of permutations of $[\![n]\!]$, the symmetric group.

Ranking Regression

Consider:

- A set of n items: $[\![n]\!] = \{1, \ldots, n\}$
- A individual expresses her preferences as (full) ranking, i.e a strict order $\succ$ over n :

$$a_1 \succ a_2 \succ \cdots \succ a_n$$

- Also seen as the permutation σ that maps an item to its rank:

$$a_1 \succ \cdots \succ a_n \quad \Leftrightarrow \quad \sigma \in \mathfrak{S}_n \text{ such that } \sigma(a_i) = i$$

$\mathfrak{S}_n$: set of permutations of $[\![n]\!]$, the symmetric group.

Problem: Given a vector X (e.g, the characteristics of an individual), the goal is to predict (her preferences) a random permutation Σ in $\mathfrak{S}_n$.

Ranking Regression

Consider:

- A set of n items: $[\![n]\!] = \{1, \ldots, n\}$
- A individual expresses her preferences as (full) ranking, i.e a strict order $\succ$ over n :

$$a_1 \succ a_2 \succ \cdots \succ a_n$$

- Also seen as the permutation σ that maps an item to its rank:

$$a_1 \succ \cdots \succ a_n \quad \Leftrightarrow \quad \sigma \in \mathfrak{S}_n \text{ such that } \sigma(a_i) = i$$

$\mathfrak{S}_n$: set of permutations of $[\![n]\!]$, the symmetric group.

Problem: Given a vector X (e.g, the characteristics of an individual), the goal is to predict (her preferences) a random permutation Σ in $\mathfrak{S}_n$.

Learn a predictive ranking rule :

$$\begin{array}{cccc} s & : & \mathcal{X} & \rightarrow & \mathfrak{S}_n \\ & & x & \mapsto & s(x) \end{array}$$

Related Work

- Has been referred to as **label ranking** in the literature [Tsoumakas et al., 2009], [Vembu and Gärtner, 2010]
- Related to multiclass and multilabel classification
- A lot of applications, e.g : document categorization, meta-learning
  - rank a set of topics relevant for a given document
  - rank a set of algorithms according to their suitability for a new dataset, based on the characteristics of the dataset
- A lot of approaches rely on parametric modelling [Cheng and Hüllermeier, 2009], [Cheng et al., 2009], [Cheng et al., 2010]

Related Work

- Has been referred to as **label ranking** in the literature [Tsoumakas et al., 2009], [Vembu and Gärtner, 2010]
- Related to multiclass and multilabel classification
- A lot of applications, e.g : document categorization, meta-learning
  - rank a set of topics relevant for a given document
  - rank a set of algorithms according to their suitability for a new dataset, based on the characteristics of the dataset
- A lot of approaches rely on parametric modelling [Cheng and Hüllermeier, 2009], [Cheng et al., 2009], [Cheng et al., 2010]

$\Rightarrow$ We develop an approach free of any parametric assumptions (**local learning**) relying on results and framework developped in [Korba et al., 2017] for **ranking aggregation**.

Outline

1. Background and Results on Ranking Aggregation
2. Ranking Median Regression
3. Local Consensus Methods for Ranking Median Regression

Outline

Background and Results on Ranking Aggregation

Ranking Median Regression

Local Consensus Methods for Ranking Median Regression

Ranking Aggregation

Suppose we have a dataset of rankings/permutations $\mathcal{D}_N = (\sigma_1, \ldots, \sigma_N) \in \mathfrak{S}_n^N$. We want to find a global order ("consensus") σ^* on the n items that best represents the dataset.

Ranking Aggregation

Suppose we have a dataset of rankings/permutations $\mathcal{D}_N = (\sigma_1, \ldots, \sigma_N) \in \mathfrak{S}_n^N$. We want to find a global order ("consensus") σ^* on the n items that best represents the dataset.

Kemeny's rule (1959)

Find the solution of :

$$\sigma^* = \underset{\sigma \in \mathfrak{S}_n}{\text{argmin}} \sum_{k=1}^{N} d(\sigma, \sigma_k)$$

where d is the Kendall's tau distance:

$$d_\tau(\sigma, \sigma') = \sum_{i<j} \mathbb{I}\{(\sigma(i) - \sigma(j))(\sigma'(i) - \sigma'(j)) < 0\},$$

Ex: σ= 1234, σ'= 2413 $\Rightarrow d_\tau(\sigma, \sigma') = 3$ (disagree on (12),(14),(34)).

Statistical Ranking Aggregation [Korba et al., 2017]

Probabilistic Modeling

$$\mathcal{D}_N = (\Sigma_1, \ldots, \Sigma_N) \quad \text{with} \quad \Sigma_k \sim P$$

where $P \sim \mathfrak{S}_n$.

Statistical Ranking Aggregation [Korba et al., 2017]

Probabilistic Modeling

$$\mathcal{D}_N = (\Sigma_1, \ldots, \Sigma_N) \qquad \text{with} \qquad \Sigma_k \sim P$$

where $P \sim \mathfrak{S}_n$.

Definition

A **Kemeny median** of P is solution of:

$$\sigma_P^* = \underset{\sigma \in \mathfrak{S}_n}{\text{argmin}}\, L_P(\sigma), \tag{1}$$

where $L_P(\sigma) = \mathbb{E}_{\Sigma \sim P}[d(\sigma, \Sigma)]$ is **the risk** of σ.

Problem: Solving (1) is NP-hard.

Exact Solutions [Korba et al., 2017]

Let $p_{i,j} = \mathbb{P}[\Sigma(i) < \Sigma(j)]$ (probability that item $i \succ j$).

Exact Solutions [Korba et al., 2017]

Let $p_{i,j} = \mathbb{P}[\Sigma(i) < \Sigma(j)]$ (probability that item $i \succ j$).

Stochastic Transitivity:

$$p_{i,j} \geq 1/2 \text{ and } p_{j,k} \geq 1/2 \Rightarrow p_{i,k} \geq 1/2.$$

Exact Solutions [Korba et al., 2017]

Let $p_{i,j} = \mathbb{P}[\Sigma(i) < \Sigma(j)]$ (probability that item $i \succ j$).

Stochastic Transitivity:

$$p_{i,j} \geq 1/2 \text{ and } p_{j,k} \geq 1/2 \Rightarrow p_{i,k} \geq 1/2.$$

In addition, if $p_{i,j} \neq 1/2$ for all $i < j$, P is said to be "strictly stochastically transitive" (**SST**)

Exact Solutions [Korba et al., 2017]

Let $p_{i,j} = \mathbb{P}[\Sigma(i) < \Sigma(j)]$ (probability that item $i \succ j$).

Stochastic Transitivity:

$$p_{i,j} \geq 1/2 \text{ and } p_{j,k} \geq 1/2 \Rightarrow p_{i,k} \geq 1/2.$$

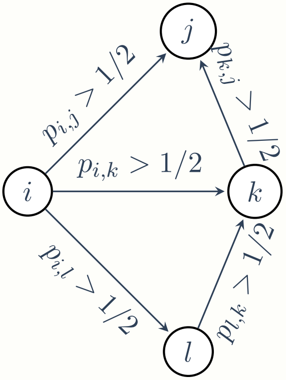
In addition, if $p_{i,j} \neq 1/2$ for all $i < j$, P is said to be ”strictly stochastically transitive” (**SST**)

Optimality

If P satisfies **SST**, its Kemeny median is **unique** and is given by its Copeland ranking:

$$\sigma^*_P(i) = 1 + \sum_{j \neq i} \mathbb{I}\{p_{i,j} < \frac{1}{2}\} \quad \text{for } 1 \leq i \leq n$$

Graph of pairwise probabilities



$$\sigma^*_P(i) = 1 + \sum_{j \neq i} \mathbb{I}\{p_{i,j} < \frac{1}{2}\}$$

$\Rightarrow$ sort the i's by increasing input degree

Exponential rates [Korba et al., 2017]

Low-Noise condition **NA**(h) for some $h > 0$:

$$\min_{i<j} |p_{i,j} - 1/2| \geq h.$$

Generalization.

Suppose P satisfies **SST and NA**(h) for a given $h > 0$. Then with overwhelming probability $1 - \frac{n(n-1)}{4} e^{-\alpha_h N}$:

Exponential rates [Korba et al., 2017]

Low-Noise condition **NA**(h) for some $h > 0$:

$$\min_{i<j} |p_{i,j} - 1/2| \geq h.$$

Generalization.

Suppose P satisfies **SST and NA**(h) for a given $h > 0$. Then with overwhelming probability $1 - \frac{n(n-1)}{4} e^{-\alpha_h N}$:

$\widehat{P}$ also verifies **SST**...

Exponential rates [Korba et al., 2017]

Low-Noise condition **NA**(h) for some $h > 0$:

$$\min_{i<j} |p_{i,j} - 1/2| \geq h.$$

Generalization.

Suppose P satisfies **SST and NA**(h) for a given $h > 0$. Then with overwhelming probability $1 - \frac{n(n-1)}{4} e^{-\alpha_h N}$:

$\widehat{P}$ also verifies **SST**...and the Kemeny median of P is given by the empirical Copeland ranking:

$$\sigma^*_P(i) = 1 + \sum_{j \neq i} \mathbb{I}\{\widehat{p}_{i,j} < \frac{1}{2}\} \quad \text{for } 1 \leq i \leq n$$

$\Rightarrow$ Under the needed conditions, empirical Copeland method

($\mathcal{O}(N\binom{n}{2})$) outputs the true Kemeny consensus (NP-hard) with high probability!

In practice: Pseudo-empirical Kemeny Medians

- If $\widehat{P}$ is SST, compute $\sigma^*_{\widehat{P}}$ with Copeland method based on $\widehat{p}_{i,j}$

In practice: Pseudo-empirical Kemeny Medians

- If $\widehat{P}$ is SST, compute $\sigma^*_{\widehat{P}}$ with Copeland method based on $\widehat{p}_{i,j}$
- Else, compute $\widetilde{\sigma}^*_{\widehat{P}}$ with empirical Borda count ([Jiang et al., 2011])

$$\widetilde{\sigma}^*_{\widehat{P}}(i) = \frac{1}{N}\sum_{k=1}^{N}\Sigma_k(i) \quad \text{for } 1 \leq i \leq n$$

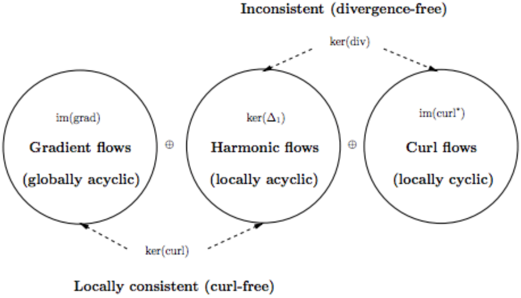


FIGURE 2. Hodge/Helmholtz decomposition of pairwise rankings

Outline

Background and Results on Ranking Aggregation

Ranking Median Regression

Local Consensus Methods for Ranking Median Regression

Our Problem

Suppose we observe $(X_1, \Sigma_1), \ldots, (X_N, \Sigma_N)$ i.i.d. copies of the pair $(X,\ \Sigma)$, where

- $X \sim \mu$, where μ is a distribution on some feature space $\mathcal{X}$
- $\Sigma \sim P_X$, where P_X is the conditional probability distribution (on $\mathfrak{S}_n$): $P_X(\sigma) = \mathbb{P}[\Sigma = \sigma | X]$

Ex: Users i with characteristics X_i order items by preference resulting in Σ_i.

Goal: Learn a predictive ranking rule :

$$\begin{array}{ccccc} s & : & \mathcal{X} & \to & \mathfrak{S}_n \\ & & x & \mapsto & s(x) \end{array}$$

which given a random vector X, predicts the permutation Σ on the n items.

Performance: Measured by the risk:

$$\mathcal{R}(s) = \mathbb{E}_{X \sim \mu, \Sigma \sim P_X} \left[d_\tau \left(s(X), \Sigma \right) \right]$$

Ranking Median Regression Approach

$$\mathcal{R}(s) = \mathbb{E}_{X\sim\mu}\left[\mathbb{E}_{\Sigma\sim P_X}\left[d_\tau\left(s(X),\Sigma\right)\right]\right] = \mathbb{E}_{X\sim\mu}\left[L_{P_X}(s(X))\right] \quad (2)$$

Ranking Median Regression Approach

$$\mathcal{R}(s) = \mathbb{E}_{X\sim\mu}\left[\mathbb{E}_{\Sigma\sim P_X}\left[d_\tau\left(s(X),\Sigma\right)\right]\right] = \mathbb{E}_{X\sim\mu}\left[L_{P_X}(s(X))\right] \quad (2)$$

Assumption

For $X \in \mathcal{X}$, P_X is **SST**: $\Rightarrow \sigma^*_{P_X} = \text{argmin}_{\sigma\in\mathfrak{S}_n} L_{P_X}(\sigma)$ is **unique**.

Optimal elements

The predictors s^* minimizing (2) are the ones that maps any point $X \in \mathcal{X}$ to the **conditional** Kemeny median of P_X:

$$s^* = \underset{s\in\mathcal{S}}{\text{argmin}}\, \mathcal{R}(s) \quad \Leftrightarrow \quad s^*(X) = \sigma^*_{P_X}$$

Ranking Median Regression Approach

$$\mathcal{R}(s) = \mathbb{E}_{X\sim\mu}\left[\mathbb{E}_{\Sigma\sim P_X}\left[d_\tau\left(s(X),\Sigma\right)\right]\right] = \mathbb{E}_{X\sim\mu}\left[L_{P_X}(s(X))\right] \quad (2)$$

Assumption

For $X \in \mathcal{X}$, P_X is **SST**: $\Rightarrow \sigma^*_{P_X} = \text{argmin}_{\sigma\in\mathfrak{S}_n} L_{P_X}(\sigma)$ is **unique**.

Optimal elements

The predictors s^* minimizing (2) are the ones that maps any point $X \in \mathcal{X}$ to the **conditional** Kemeny median of P_X:

$$s^* = \underset{s\in\mathcal{S}}{\text{argmin}}\, \mathcal{R}(s) \quad \Leftrightarrow \quad s^*(X) = \sigma^*_{P_X}$$

To minimize (2) approximately:

$$\sigma^*_{P_X} \text{ for any } X \Longrightarrow \sigma^*_{P_{\mathcal{C}}} \text{ for any } X \in \mathcal{C}$$

$\Rightarrow$ We develop Local consensus methods.

Statistical Framework- ERM

Optimize a statistical version of the theoretical risk based on the training data (X_k, Σ_k)'s:

$$\min_{s \in \mathcal{S}} \widehat{\mathcal{R}}_N(s) = \frac{1}{N} \sum_{k=1}^{N} d_\tau(s(X_k),\ \Sigma_k)$$

where $\mathcal{S}$ is the set of measurable mappings.

Statistical Framework- ERM

Optimize a statistical version of the theoretical risk based on the training data (X_k, Σ_k)'s:

$$\min_{s \in \mathcal{S}} \widehat{\mathcal{R}}_N(s) = \frac{1}{N} \sum_{k=1}^{N} d_\tau(s(X_k),\ \Sigma_k)$$

where $\mathcal{S}$ is the set of measurable mappings.

$\Rightarrow$ We will consider a subset $\mathcal{S}_{\mathcal{P}} \subset \mathcal{S}$:

- rich enough so that $\inf_{s \in \mathcal{S}_{\mathcal{P}}} \mathcal{R}(s) - \inf_{s \in \mathcal{S}} \mathcal{R}(s)$ is "small"
- ideally appropriate for greedy optimization.

$\Rightarrow$ $\mathcal{S}_{\mathcal{P}}$= space of piecewise constant ranking rules ("local consensus methods")

Outline

Background and Results on Ranking Aggregation

Ranking Median Regression

Local Consensus Methods for Ranking Median Regression

Piecewise Constant Ranking Rules

Let $\mathcal{P} = \{\mathcal{C}_1, \ \ldots, \ \mathcal{C}_K\}$ be a partition of the feature space $\mathcal{X}$.

Any $s \in \mathcal{S}_{\mathcal{P}}$ (ranking rules that are constant on each cell of $\mathcal{P}$) can be written as:

$$s_{\mathcal{P},\bar{\sigma}}(x) = \sum_{k=1}^{K} \sigma_k \cdot \mathbb{I}\{x \in \mathcal{C}_k\} \text{ where } \bar{\sigma} = (\sigma_1, \ldots, \sigma_K)$$

Piecewise Constant Ranking Rules

Let $\mathcal{P} = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ be a partition of the feature space $\mathcal{X}$.

Any $s \in \mathcal{S}_{\mathcal{P}}$ (ranking rules that are constant on each cell of $\mathcal{P}$) can be written as:

$$s_{\mathcal{P},\bar{\sigma}}(x) = \sum_{k=1}^{K} \sigma_k \cdot \mathbb{I}\{x \in \mathcal{C}_k\} \text{ where } \bar{\sigma} = (\sigma_1, \ldots, \sigma_K)$$

Local Learning

Let $P_{\mathcal{C}_k}$ the cond. distr. of Σ given $X \in \mathcal{C}_k$:
$P_{\mathcal{C}_k}(\sigma) = \mathbb{P}[\Sigma = \sigma | X \in \mathcal{C}_k]$

Piecewise Constant Ranking Rules

Let $\mathcal{P} = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ be a partition of the feature space $\mathcal{X}$.

Any $s \in \mathcal{S}_{\mathcal{P}}$ (ranking rules that are constant on each cell of $\mathcal{P}$) can be written as:

$$s_{\mathcal{P},\bar{\sigma}}(x) = \sum_{k=1}^{K} \sigma_k \cdot \mathbb{I}\{x \in \mathcal{C}_k\} \text{ where } \bar{\sigma} = (\sigma_1, \ldots, \sigma_K)$$

Local Learning

Let $P_{\mathcal{C}_k}$ the cond. distr. of Σ given $X \in \mathcal{C}_k$:
$P_{\mathcal{C}_k}(\sigma) = \mathbb{P}[\Sigma = \sigma | X \in \mathcal{C}_k]$

Recall: P_X is SST for any $X \in \mathcal{X}$.

Idea: $P_{\mathcal{C}_k}$ is still SST and $\sigma^*_{P_{\mathcal{C}}} = \sigma^*_{P_X}$ provided the $\mathcal{C}_k$'s are small enough.

Theorem

Suppose that:

There exists $M < \infty$ such that:

$\forall (x, x') \in \mathcal{X}^2, \;\; \sum_{i<j} |p_{i,j}(x) - p_{i,j}(x')| \leq M \cdot ||x - x'||$.

Then:

$$\mathcal{R}(s^*_{\mathcal{P}}) - \mathcal{R}(s^*) \leq M.\delta_{\mathcal{P}}$$

where $\delta_{\mathcal{P}}$ is the max. diameter of $\mathcal{P}$'s cells.

Theorem

Suppose that:

There exists $M < \infty$ such that:

$\forall (x, x') \in \mathcal{X}^2, \quad \sum_{i<j} |p_{i,j}(x) - p_{i,j}(x')| \leq M \cdot ||x - x'||$.

Then:

$$\mathcal{R}(s^*_{\mathcal{P}}) - \mathcal{R}(s^*) \leq M.\delta_{\mathcal{P}}$$

where $\delta_{\mathcal{P}}$ is the max. diameter of $\mathcal{P}$'s cells.

Suppose in addition that:

For all $x \in \mathcal{X}$, $P_x \in \mathcal{T}$ and $H = \inf_{x \in \mathcal{X}} \min_{i<j} |p_{i,j}(x) - 1/2| > 0$.

and that $P_{\mathcal{C}} \in \mathcal{T}$ for all $\mathcal{C} \in \mathcal{P}$.

Then,

$$\mathbb{E}\left[d_\tau\left(\sigma^*_{P_X}, s^*_{\mathcal{P}}(X)\right)\right] \leq \sup_{x \in \mathcal{X}} d_\tau\left(\sigma^*_{P_x}, s^*_{\mathcal{P}}(x)\right) \leq (M/H) \cdot \delta_{\mathcal{P}}$$

Partitioning Methods

Goal: Generate partitions $\mathcal{P}_N$ in a data-driven fashion.

For $\mathcal{C} \in \mathcal{P}_N$, consider its empirical distribution:

$$\widehat{P}_{\mathcal{C}} = \frac{1}{N_{\mathcal{C}}} \sum_{k: X_k \in \mathcal{C}} \delta_{\Sigma_k}$$

and compute locally its Pseudo-Empirical Kemeny median $\widetilde{\sigma}^*_{\widehat{P}_{\mathcal{C}}}$.

Partitioning Methods

Goal: Generate partitions $\mathcal{P}_N$ in a data-driven fashion.

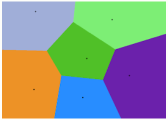
For $\mathcal{C} \in \mathcal{P}_N$, consider its empirical distribution:

$$\widehat{P}_{\mathcal{C}} = \frac{1}{N_{\mathcal{C}}} \sum_{k: X_k \in \mathcal{C}} \delta_{\Sigma_k}$$

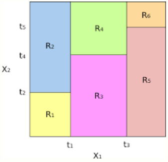
and compute locally its Pseudo-Empirical Kemeny median $\widetilde{\sigma}^*_{\widehat{P}_{\mathcal{C}}}$.

Two methods are investigated:

- k-nearest neighbor (Voronoi partitioning)



- decision tree (Recursive partitioning)



K-Nearest Neigbors Algorithm

1. Fix $k \in \{1, \ldots, N\}$ and a query point $x \in \mathcal{X}$
2. Sort $(X_1, \Sigma_1), \ldots, (X_N, \Sigma_N)$ by increasing order of the distance to x : $\|X_{(1,N)} - x\| \leq \ldots \leq \|X_{(N,N)} - x\|$
3. Consider next the empirical distribution calculated using the k training points closest to x

$$\widehat{P}(x) = \frac{1}{k} \sum_{l=1}^{k} \delta_{\Sigma_{(l,N)}}$$

   and compute the pseudo-empirical Kemeny median, yielding the k-NN prediction at x:

$$s_{k,N}(x) \overset{def}{=} \widetilde{\sigma}^*_{\widehat{P}(x)}.$$

$\Rightarrow$ We recover the classical bound $\mathcal{R}(s_{k,N}) - \mathcal{R}^* = \mathcal{O}(\frac{1}{\sqrt{k}} + \frac{k}{N})$

Decision Tree

Split recursively the feature space $\mathcal{X}$ to minimize some impurity criterion.

Analog to Gini criterion in classification: m classes, f_i proportion of class $i \to I_G(f) = \sum_{i=1}^{m} f_i(1 - f_i)$

Here, for a cell $\mathcal{C} \in \mathcal{P}_N$:

- Impurity [Alvo and Philip, 2014]:

$$\gamma_{\widehat{P}_{\mathcal{C}}} = \frac{1}{2} \sum_{i<j} \widehat{p}_{i,j}(\mathcal{C}) \left(1 - \widehat{p}_{i,j}(\mathcal{C})\right)$$

  which is tractable and satisfies the double inequality

$$\widehat{\gamma}_{\widehat{P}_{\mathcal{C}}} \leq \min_{\sigma \in \mathfrak{S}_n} L_{\widehat{P}_{\mathcal{C}}}(\sigma) \leq 2\widehat{\gamma}_{\widehat{P}_{\mathcal{C}}}.$$

- Terminal value : Compute the pseudo-empirical median of a cell $\mathcal{C}$:

$$s_{\mathcal{C}}(x) \overset{def}{=} \widetilde{\sigma}^*_{\widehat{P}_{\mathcal{C}}}.$$

Conclusion

Interesting challenges:

- Most of the maths from euclidean spaces cannot be applied
- But our insights still hold
- Based on our results on ranking aggregation, we develop a novel approach to ranking regression/label ranking
- Theoretical guarantees (approximation error, rates of convergence)
- We propose two practical algorithms

Openings: How to extend to incomplete rankings (+with ties)?

Alvo, M. and Philip, L. (2014).
Decision tree models for ranking data.
In *Statistical Methods for Ranking Data*, pages 199–222.
Springer.

Cheng, W., Dembczyński, K., and Hüllermeier, E. (2010).
Label ranking methods based on the Plackett-Luce model.
In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 215–222.

Cheng, W., Hühn, J., and Hüllermeier, E. (2009).
Decision tree and instance-based learning for label ranking.
In *Proceedings of the 26th International Conference on Machine Learning (ICML-09)*, pages 161–168.

Cheng, W. and Hüllermeier, E. (2009).
A new instance-based label ranking approach using the mallows model.
Advances in Neural Networks–ISNN 2009, pages 707–716.

Jiang, X., Lim, L.-H., Yao, Y., and Ye, Y. (2011).
Statistical ranking and combinatorial hodge theory.
Mathematical Programming, 127(1):203–244.

Korba, A., Clémençon, S., and Sibony, E. (2017).
A learning theory of ranking aggregation.
In *Proceeding of AISTATS 2017*.

Tsoumakas, G., Katakis, I., and Vlahavas, I. (2009).
Mining multi-label data.
In *Data mining and knowledge discovery handbook*, pages 667–685. Springer.

Vembu, S. and Gärtner, T. (2010).
Label ranking algorithms: A survey.
In *Preference learning*, pages 45–64. Springer.