A Learning Theory of Ranking Aggregation

Anna Korba*, Stephan Clémençon* and Eric Sibony†

* LTCI, Télécom ParisTech, Université Paris-Saclay , † Shift Technology

Ranking Aggregation

In the simplest formulation, a full ranking on a set of items $[\![n]\!]$ is seen as the permutation $\sigma \in \mathfrak{S}_n$ that maps an item i to its rank $\sigma(i)$. Given a collection of $N \geq 1$ permutations $\sigma_1, \ldots, \sigma_N$, the goal of ranking aggregation is to find $\sigma^* \in \mathfrak{S}_n$ that best summarizes it. A popular approach, called **Kemeny's rule**, consists in solving the **NP-hard** following optimization problem:

$$\min_{\sigma \in \mathfrak{S}_n} \sum_{t=1}^{N} d(\sigma_t, \sigma),$$

where $d(.,\ .)$ is the Kendall's tau distance, i.e.:

$$d(\sigma, \sigma') = \sum_{1 \leq i < j \leq n} \mathbb{I}\{(\sigma(j) - \sigma(i))(\sigma'(j) - \sigma'(i)) < 0\}$$

Previous work: Numerous results, e.g. bounds on the cost of approximation procedures, consistency relationships between Kemeny aggregation and other voting rules...

Our contribution: In a general statistical framework for Kemeny aggregation, we describe optimal elements and provide statistical guarantees for the generalization properties of an empirical median ranking in the form of rate bounds.

Statistical Framework

Suppose that the dataset is composed of N i.i.d copies $\Sigma_1, \ldots, \Sigma_N$ of a generic random variable $\Sigma \sim P$. A true median for P w.r.t d is any solution of the minimization problem:

$$\min_{\sigma \in \mathfrak{S}_n} L(\sigma),$$

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

What is the performance of **Kemeny empirical medians**, i.e. solutions $\widehat{\sigma}_N$ of

$$\min_{\sigma \in \mathfrak{S}_n} \widehat{L}_N(\sigma), \qquad (1)$$

where $\widehat{L}_N(\sigma) = \frac{1}{N}\sum_{t=1}^{N} d(\Sigma_t, \sigma)$?

Optimality of a Consensus

The risk of a permutation candidate $\sigma \in \mathfrak{S}_n$ can be written as

$$L(\sigma) = \sum_{i<j} p_{i,j}\mathbb{I}\{\sigma(i) > \sigma(j)\} + \sum_{i<j}(1 - p_{i,j})\mathbb{I}\{\sigma(i) < \sigma(j)\}.$$

So if $\exists$ a permutation σ with the property that $\forall i < j$ s.t. $p_{i,j} \neq 1/2$,

$$(\sigma(j) - \sigma(i)) \cdot (p_{i,j} - 1/2) > 0, \qquad (2)$$

it would be necessarily a median for P.

Definition 1. *The probability distribution P on $\mathfrak{S}_n$ is said to be stochastically transitive if it fulfills the condition: $\forall (i,j,k) \in [\![n]\!]^3$,*

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

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

Theorem 1. *If the distribution P is stochastically transitive, there exists $\sigma^* \in \mathfrak{S}_n$ such that (2) holds true. In this case, we have*

$$L^* = \sum_{i<j} \left\{ \frac{1}{2} - \left| p_{i,j} - \frac{1}{2} \right| \right\},$$

the excess of risk of any $\sigma \in \mathfrak{S}_n$ is given by

$$L(\sigma) - L^* = 2\sum_{i<j} |p_{i,j} - 1/2| \cdot \mathbb{I}\{(\sigma(j) - \sigma(i))(p_{i,j} - 1/2) < 0\}$$

and the set of medians of P is the class of equivalence of σ^* w.r.t. the equivalence relationship:

$$\sigma \mathcal{R}_P \sigma' \Leftrightarrow (\sigma(j) - \sigma(i))(\sigma'(j) - \sigma'(i)) > 0 \quad \text{for all } i < j \text{ such that } p_{i,j} \neq 1/2.$$

*In addition, the mapping s^* (equivalent of the **Copeland score**) defined by*

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

belongs to $\mathfrak{S}_n$ and is the unique median of P iff P is strictly stochastically positive.

Connection to Other Voting Rules

Extension of voting rules to a distribution P:

- Copeland score of item i: $s(i) = \sum_{k \neq i} \mathbb{I}\{p_{i,k} \leq 1/2\} - \mathbb{I}\{p_{i,k} > 1/2\}$
- Borda score of item i: $s(i) = \mathbb{E}_P[\Sigma(i)]$

Proposition 1. (Borda consensus) *The probability distribution P on $\mathfrak{S}_n$ is said to be strongly stochastically transitive if $\forall (i,j,k) \in [\![n]\!]^3$:*

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

Then under this condition, and for $i < j$, $p_{i,j} \neq \frac{1}{2}$, there exists a unique $\sigma^* \in \mathfrak{S}_n$ such that (2) holds true, corresponding to the Kemeny and Borda consensus both at the same time.

Universal Rates

The performance of a Kemeny empirical median $\widehat{\sigma}_N$ is mesured by its excess risk:

$$L(\widehat{\sigma}_N) - L^* \leq 2 \max_{\sigma \in \mathfrak{S}_n} |\widehat{L}_N(\sigma) - L(\sigma)|.$$

We establish the following result.

Proposition 2. *The excess risk of $\widehat{\sigma}_N$ is upper bounded:*

(i) In expectation by

$$\mathbb{E}\left[L(\widehat{\sigma}_N) - L^*\right] \leq \frac{n(n-1)}{2\sqrt{N}}$$

(ii) With probability higher than $1 - \delta$ for any $\delta \in (0,1)$ by

$$L(\widehat{\sigma}_N) - L^* \leq \frac{n(n-1)}{2}\sqrt{\frac{2\log(n(n-1)/\delta)}{N}}.$$

We then establish the tightness of the upper bound by providing a lower bound for the **minimax risk** :

$$\mathcal{R}_N \overset{def}{=} \inf_{\sigma_N} \sup_{P} \mathbb{E}_P\left[L_P(\sigma_N) - L_P^*\right], \qquad (3)$$

where the sup. is taken over all distr. on $\mathfrak{S}_n$.

Proposition 3. *The minimax risk for Kemeny aggregation is lower bounded as follows:*

$$\mathcal{R}_N \geq \frac{1}{16e\sqrt{N}}.$$

Fast Rates

For $h > 0$, we define the low noise condition:

$$\mathbf{NA}(h)\text{: } \min_{i<j} |p_{i,j} - 1/2| \geq h.$$

Let $\widehat{p}_{i,j} = (1/N)\sum_{m=1}^{N} \mathbb{I}\{\Sigma_m(i) < \Sigma_m(j)\}$. We establish exponential rates of convergence.

Proposition 4. *Assume that P is stochastically transitive and fulfills condition $\mathbf{NA}(h)$ for some $h > 0$. The following assertions hold true.*

(i) For any empirical Kemeny median $\widehat{\sigma}_N$, we have:

$$\mathbb{E}\left[L(\widehat{\sigma}_N) - L^*\right] \leq \frac{n^2(n-1)^2}{8} e^{-\frac{N}{2}\log\left(\frac{1}{1-4h^2}\right)}.$$

(ii) With probability at least $1 - (n(n-1)/4)e^{-\frac{N}{2}\log\left(\frac{1}{1-4h^2}\right)}$, the mapping

$$\widehat{s}_N(i) = 1 + \sum_{k \neq i} \mathbb{I}\{\widehat{p}_{i,k} < \frac{1}{2}\}$$

for $1 \leq i \leq n$ belongs to $\mathfrak{S}_n$ and is the unique solution of (1).

Proposition 5. *Let $h > 0$ and define*

$$\widetilde{\mathcal{R}}_N(h) = \inf_{\sigma_N} \sup_{P} \mathbb{E}_P\left[L_P(\sigma_N) - L_P^*\right],$$

where the sup. is taken over all stochastically transitive distr. on $\mathfrak{S}_n$ satisfying $\mathbf{NA}(h)$. We have:

$$\widetilde{\mathcal{R}}_N(h) \geq \frac{h}{4} e^{-N2h\log\left(\frac{1+2h}{1-2h}\right)}. \qquad (4)$$

Let $\alpha_h = \frac{1}{2}log\left(1/(1-4h^2)\right)$ and $\beta_h = 2h\log((1+2h)/(1-2h))$. We have $\alpha_h \sim \frac{1}{2}\beta_h$ when $h \to \frac{1}{2}$.

Computational Benefit

Under the low-noise condition, the Copeland method (complexity $O(N\binom{n}{2})$) outputs the exact NP-hard Kemeny consensus (Proposition 4 (ii)).

References

[1] J.Y. Audibert and A.B. Tsybakov. *Fast Learning Rates For Plug-in Classifiers*. Annals of Statistics, 2007.

[2] V. Koltchinskii and O. Beznosova. *Exponential Convergence Rates in Classification*. COLT, 2005.