Judging competitions and benchmarks: a candidate election approach SUPPLEMENTAL MATERIAL

Contents

1 Theoretical results

All the theoretical results presented in the paper were found in the literature, except for *Success rate* and *Relative Difference*. Let's find out by ourselves.

1.1 Criteria satisfied by success rate method

(1) Majority: No, *Success rate* doesn't satisfy majority criterion. Counter-example:

	j_1	j_2	j_3
A	1	1	0
B	0.8	0.8	1
C	0.6	0.6	0.6

Table 1: This is a score matrix exposing a counter example. A is ranked first by a majority of judges but its average success rate is the same as B: $\frac{2}{3}$.

(2) Condorcet: No, *Success rate* doesn't satisfy Condorcet criterion, as implied by the fact it does not satisfy majority criterion (Majority $\in$ Condorcet).

(3) Consistency: Yes, *Success rate* meets consistency criterion and **(4) participation criterion**. A ranked system is consistent and meets participation criterion if it's a scoring function (i.e. positional system) [9]. Success rate is positional as adding a judge improve the score of the candidates according the their position in the ranking of this judge.

(5) LIIA: No, *Success rate* is not LIIA. Counter example:

	j_1	j_2	j_3
A	0.6	0.6	0
B	0.4	0.4	1
C	1	0	0.4

Table 2: This is a score matrix exposing a counter example. If there is only A and B then A wins. If we add C then it is a tie.

(6) IIA: No, *Success rate* is not IIA, as implied by the fact it is not LIIA (LIIA $\in$ IIA).

(7) Clone-proof: No, *Success rate* is not clone-proof. Counter-example:

	A	B	C	Average
A	-	0.6	0.4	0.5
B	0.4	-	0.5	0.45
C	0.6	0.5	-	0.55

Table 3: This is a pairwise success rate table exposing a counter example. If we repeatedly duplicate the candidate B, A will end up in front of C.

1.2 Criteria satisfied by relative difference method

(1) Majority: No, *Relative Difference* doesn't satisfy majority criterion. Counter example found empirically (Majority rate $\neq 1$).

(2) Condorcet: No, *Relative Difference* doesn't satisfy Condorcet criterion. Counter example found empirically (Condorcet rate $\neq 1$). It's also implied by the fact that this method does not satisfy majority criterion (Majority $\in$ Condorcet).

(3) Consistency: Yes, *Relative Difference* meets consistency.

Let's suppose that:

1. $\texttt{rank}\Big(f(X)\Big) = \texttt{rank}\Big(f(Y)\Big)$
2. $\texttt{rank}\Big(f([X\ Y])\Big) \neq \texttt{rank}\Big(f(X)\Big)$

Without loss of generality, let's assume that the candidate ranked first is not the same in $\texttt{rank}\Big(f(X)\Big)$ and $\texttt{rank}\Big(f([X\ Y])\Big)$, respectively at index α and β.

Then:

$$\texttt{rank}\Big(f([X\ Y])\Big)_\alpha > \texttt{rank}\Big(f([X\ Y])\Big)_\beta$$

$$\implies f([X\ Y])_\alpha < f([X\ Y])_\beta$$

From the definition of f:

$$\sum_{j\neq\alpha} \frac{1}{m} \sum_{k=1}^{m} Q_k^{\alpha,j} < \sum_{j\neq\beta} \frac{1}{m} \sum_{k=1}^{m} Q_k^{\beta,j}$$

where $Q_k^{i,j} = \frac{c_{ik}-c_{jk}}{c_{ik}+c_{jk}}$

$$\implies \frac{1}{m} \sum_{k=1}^{m} \widetilde{Q}_k^{\alpha} < \frac{1}{m} \sum_{k=1}^{m} \widetilde{Q}_k^{\beta}$$

where $\widetilde{Q}_k^i = \sum_{j\neq i} Q_k^{i,j}$

$$\implies \sum_{k=1}^{m} \widetilde{Q}_k^{\alpha} < \sum_{k=1}^{m} \widetilde{Q}_k^{\beta}$$

$$\sum_{k\in\mathcal{J}_X} \widetilde{Q}_k^{\alpha} + \sum_{k\in\mathcal{J}_Y} \widetilde{Q}_k^{\alpha} < \sum_{k\in\mathcal{J}_X} \widetilde{Q}_k^{\beta} + \sum_{k\in\mathcal{J}_Y} \widetilde{Q}_k^{\beta}$$

This is **impossible** because $\texttt{rank}\Big(f(X)\Big) = \texttt{rank}\Big(f(Y)\Big)$ which implies that $\sum_{k\in\mathcal{J}_X} \widetilde{Q}_k^{\alpha} > \sum_{k\in\mathcal{J}_X} \widetilde{Q}_k^{\beta}$ and $\sum_{k\in\mathcal{J}_Y} \widetilde{Q}_k^{\alpha} > \sum_{k\in\mathcal{J}_Y} \widetilde{Q}_k^{\beta}$ (and $\forall a, b, c, d \in \mathbb{R}$, if $a > c$ and $b > d$, then $(a + b) > (c + d)$).

In conclusion, if $\texttt{rank}\Big(f(X)\Big) = \texttt{rank}\Big(f(Y)\Big)$, then $\texttt{rank}\Big(f([X\ Y])\Big) = \texttt{rank}\Big(f(X)\Big) = \texttt{rank}\Big(f(Y)\Big)$

(4) Participation: Yes, *Relative Difference* meets participation criterion. The final score given to a candidate can be expressed as the mean of the scores that this candidate obtained on all judges. Suppose we have a judge $\mathbf{j}$ and a score matrix X, with $f(X)_u > f(X)_v$, and $j_u > j_v$.

By following the same reasoning as for consistency, we have:

$$f([X\ \mathbf{j}]) = \sum_{k\in\mathcal{J}_X} \widetilde{Q}_k + \widetilde{Q}_j$$

Also, we have:

$$\texttt{rank}\Big(f(\mathbf{j})\Big) = \texttt{rank}(\mathbf{j})$$

$$\implies \texttt{rank}\Big(\widetilde{Q}_j\Big) = \texttt{rank}(\mathbf{j})$$

We can deduce:

$$f([X\ \mathbf{j}])_u > f([X\ \mathbf{j}])_v$$

Therefore, a judge which prefers a candidate $\mathbf{u}$ against another candidate $\mathbf{v}$ can only improve the position of $\mathbf{u}$ relatively to $\mathbf{v}$.

(5) LIIA: No, *Relative Difference* is not LIIA. Counter example (same as for *Success rate* above):

	j_1	j_2	j_3
A	0.6	0.6	0
B	0.4	0.4	1
C	1	0	0.4

Table 4: This is a score matrix exposing a counter example. If there is only A and B then A wins. If we add C then it is a tie.

(6) IIA: No, *Relative Difference* is not IIA, as implied by the fact it is not LIIA (LIIA $\in$ IIA).

(7) Clone-proof: No, *Relative Difference* is not clone-proof. Counter example:

	j_1	j_2	j_3
A	0.6	0.6	0.4
B	0.5	0.4	0.6
C	0.5	0.7	0.5

Table 5: This is a score matrix exposing a counter example. If we repeatedly duplicate the candidate B, A will end up in front of C using the relative difference method.

2 Datasets-Algorithms matrices

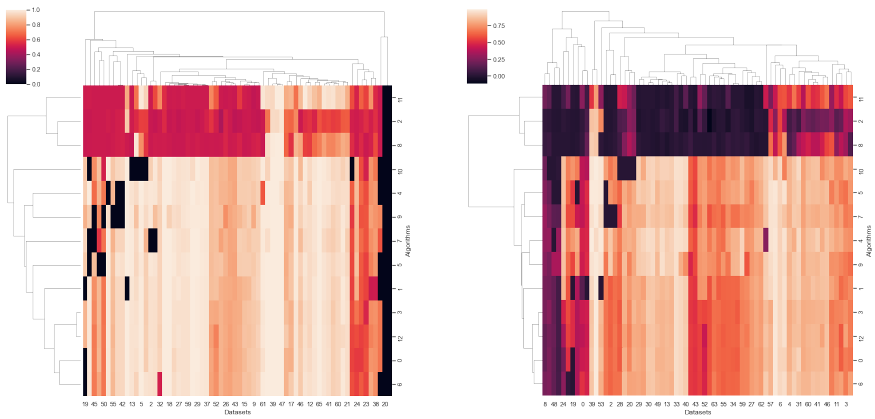

Fig. 1: Heatmap with hierarchical clustering, AutoDL-AUC (left) and AutoDL-ALC (right) score matrix.

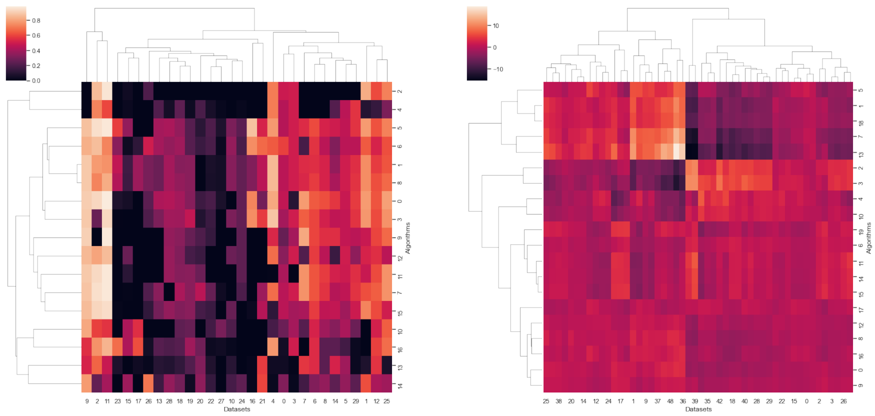

Fig. 2: Heatmap with hierarchical clustering, AutoML (left) and Artificial (right) score matrix.

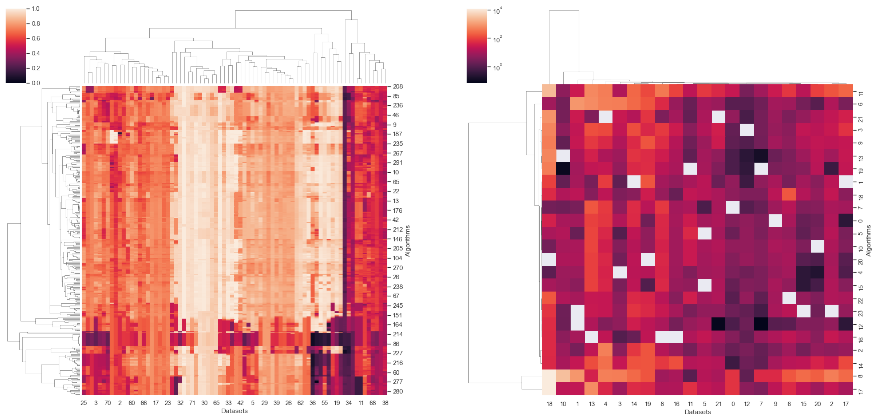

Fig. 3: Heatmap with hierarchical clustering, OpenML (left) and Statlog (right) score matrix.

3 Empirical results and interesting plots

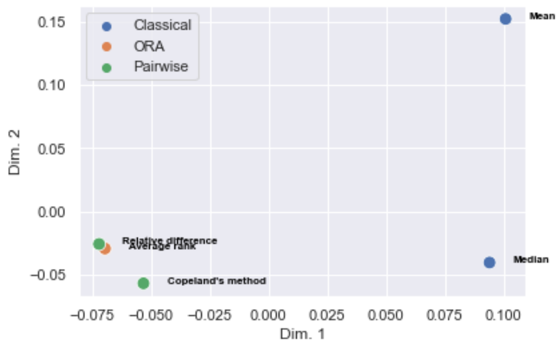

Fig. 4: Multidimensional scale (MDS) plot of the rankings produced by each ranking method. The metric used for the MDS is the Spearman distance, averaged on all benchmarks. This gives an idea of the similarities between the methods.

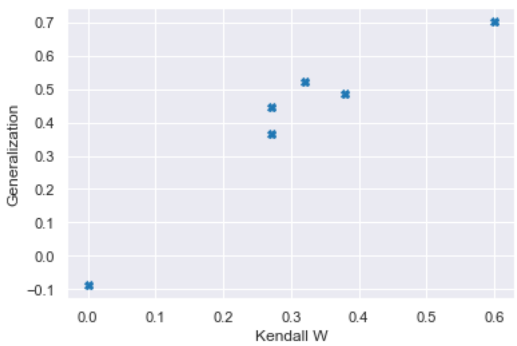

Fig. 5: Concordance of the DA matrices versus the mean generalization score obtained by the ranking functions. Each point is a benchmark.

The results on stability are summarized in Figure 6.

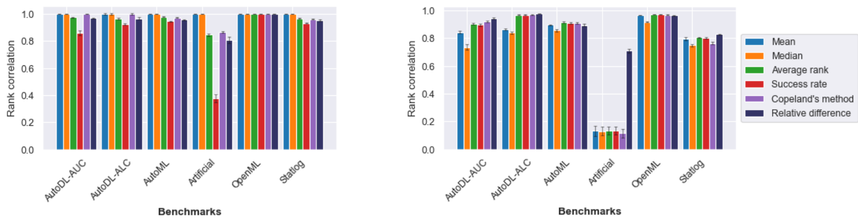

Fig. 6: Stability against candidate axis perturbation (left) and judge axis perturbation (right) of the ranking functions on each benchmark. The error bars represents the standard deviation across the scores obtained on all 10,000 repeat trials.

4 Correlation and concordance measures

4.1 Kendall's rank correlation coefficient

Intuitively, Kendall's τ measures a correlation linked to the number of "neighbor swaps" needed to transform j into j'. It's definition involves all possible pair of observations (j_c, j'_c) and $(j_{c'}, j'_{c'})$. The two pairs are said to be *concordant* if both judges j and j' agrees on their order of the candidates c and c', otherwise they are *discordant*. Kendall's τ is defined as following [5]:

$$\tau = \frac{n_c - n_d}{\binom{n}{2}}$$

where n_c and n_d represent the number of concordant pairs and discordant pairs respectively.

The actual version we use is the τ_b which accounts for ties and have a more complex definition [1]:

$$\tau_b = \frac{n_c - n_d}{\sqrt{((\binom{n}{2} - n_1)((\binom{n}{2} - n_2)}}$$

where $n_1 = \sum_i \frac{t_i(t_i-1)}{2}$ and $n_2 = \sum_{i'} \frac{u_{i'}(u_{i'}-1)}{2}$ with t_i the number of tied values in the i^{th} group of ties for the first quantity and $u_{i'}$ the number of tied values in the j^{th} group of ties for the second quantity.

In practice, Spearman's ρ and Kendall's τ seem to be correlated.

4.2 Kendall's W

To compute the concordance, we proposed to compute the mean Spearman's ρ of all possible pairs of rankings. However, this algorithm's complexity is exponential $O(2^n)$ with the number of judges. To avoid the problem of complexity, we can compute the concordance using Kendall's W statistics [4] (in practice we use a newer version of Kendall's W accounting for ties [7]).

$$W(M) = \frac{12 \sum_{i=1}^{n} (R_i - \bar{R})^2}{m^2(n^3 - n)}$$

With R_i being the total rank of candidate i on all judges:

$$R_i = \sum_{j=1}^{m} r_{i,j}$$

And $\bar{R}$ being the mean value of all total ranks:

$$\bar{R} = \frac{1}{n} \sum_{i=1}^{n} R_i$$

It has been shown in [3] that W is linearly related to $\bar{r_s}$, the mean value of the Spearman's rank correlation coefficients between all $\binom{m}{2}$ possible pairs of rankings between judges. Here is the relation between Kendall's W and the average Spearman's ρ between all possible pairs of judges:

$$\bar{r_s} = \frac{mW - 1}{m - 1}$$

We have published our implementation of Kendall's W and its second version accounting for ties in the Python Package Ranky [6], dedicated to ranking methods and measures.

5 Optimal rank aggregation and Kemeny-Young methods

Optimal rank aggregation (ORA) methods is a family of ranking methods that consists in proposing a distance function $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$ and finding a ranking r which minimizes the following objective function:

$$l(\mathbf{r}) = \sum_{\mathbf{j} \in \mathcal{J}} d(\mathbf{r}, \mathbf{j})$$

Some well-known distance functions that can be used are Kendall's τ distance, Spearman distance, Cayley distance or the Euclidean distance.

The ORA using Kendall's τ as a distance function is known as the *Kemeny-Young* method. It has interesting properties such as being a Condorcet method and satisfying Local IIA; however, its computation is NP-Hard.

In practice we perform the optimization using differential evolution [8]. A good overview of ORA and rank distance functions is given in [2]. The high complexity of *Kemeny-Young* method prevented us from including it in the experiments.

References

[1] A. Agresti. *Analysis of Ordinal Categorical Data.* New York: John Wiley & Sons, 2010.

[2] W. J. Heiser and A. D'Ambrosio. Clustering and prediction of rankings within a kemeny distance framework. In B. Lausen, D. V. den Poel, and A. Ultsch, editors, *Algorithms from and for Nature and Life - Classification and Data Analysis*, pages 19–31. Springer, 2013.

[3] M. G. Kendall and J. D. Gibbons. Rank correlation methods. *New York, NY : Oxford University Press*, 1990.

[4] M. G. Kendall and B. B. Smith. The Problem of m Rankings. *The Annals of Mathematical Statistics*, 10(3):275 – 287, 1939.

[5] R. Nelsen. Kendall tau metric. In *Encyclopedia of Mathematics*. EMS Press, 2001.

[6] A. Pavao. ranky. `https://github.com/didayolo/ranky`, 2020.

[7] S. Siegel and J. Castellan, N. John. *Nonparametric Statistics for the Behavioral Sciences (2nd ed.)*. New York: McGraw-Hill, 1988.

[8] R. Storn and K. V. Price. Differential evolution - A simple and efficient heuristic for global optimization over continuous spaces. *J. Glob. Optim.*, 11(4):341–359, 1997.

[9] H. P. Young. Social choice scoring functions. *SIAM Journal on Applied Mathematics Vol. 28, No. 4*, pages 824 – 838, 1975.