Ranking Scheme for MSD 2018

(Lena Maier-Hein, Björn Menze, Annika Reinke, Annette Kopp-Schneider)

There will be one ranking per metric (DSC: Dice Similarity Coefficient; NSD: Normalized Surface Distance) separately for both the training tasks and the mystery tasks:

1. A so-called *significance score* is determined for each algorithm \(a\) separately for each task (i.e. sub-challenge) \(c_i\) and metric \(m_j \in \{\text{DSC}, \text{NSD}\}\) and referred to as \(s_{ij}(a)\). It is computed as follows:
   - **Performance assessment per case:** Determine performance \(m_j(a_l, t_{ik})\) of all algorithms \(a_l, (l = 1, \ldots, N_A)\) for all test cases \(t_{ik}, (k = 1, \ldots, N_t)\) where \(N_t\) is the number of test cases in competition \(c_i\) and \(N_A\) is the number of competing algorithms.
     - In case of N/A value: set \(m_j(a_l, t_{ik})\) to worst possible value (i.e. 0 for DSC and NSD).
   - **Statistical tests:** perform all pairwise comparisons between algorithms \((a_l, a_{l'})\) with the values \(m_j(a_l, t_{ik}) - m_j(a_{l'}, t_{ik})\) \((k = 1, \ldots, N_t)\) using Wilcoxon signed rank test.
   - **Significance scoring:** \(s_{ij}(a_l)\) then equals the number of algorithms performing significantly worse than \(a_l\) according to the test \((\alpha = 0.05)\).
   - **Significance ranking:** The ranking is computed from the scores \(s_{ij}(a_l)\) (shared places possible) with the highest score corresponding to the best algorithm(s) (rank 1).
     - If a task has multiple sub-tasks, the ranking scheme is applied separately to each sub-task, and the final ranking is computed from the mean significance ranks.

2. The final ranking for the training tasks and mystery tasks is computed from the mean significance ranks of all algorithms, where the mean is determined from the seven training tasks and three mystery tasks, respectively.

Further reading: