Pairwise statistical comparisons of multiple algorithms

Bin-Bin JIA, Jun-Ying LIU, Min-Ling ZHANG

Frontiers of Computer Science, DOI: 10.1007/s11704-025-41325-0

Problems & Ideas

- Problems of current pairwise statistical comparisons:
  - The commonly-used average rank-based strategy (e.g., the combination of the Friedman test and the Nemenyi post-hoc test) can lead to test results that are inconsistent with common sense.
  - The Friedman test plot is not suitable to present the pairwise statistical comparative results of non-average rank-based strategies (e.g., win/tie/loss obtained by Wilcoxon signed-ranks test).
- Ideas: Use non-average rank-based strategies (e.g., Wilcoxon signed-ranks test) to conduct pairwise statistical comparisons of multiple algorithms and present the test results with an equilateral polygon.

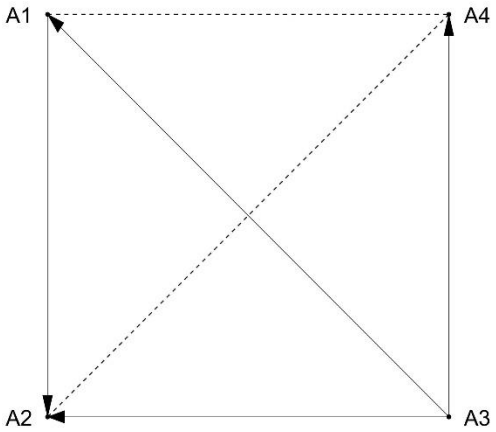
k	5	6	7	8	9	10
N ($\alpha = 0.05$)	38	57	82	111	145	184
N ($\alpha = 0.1$)	31	47	68	93	123	157

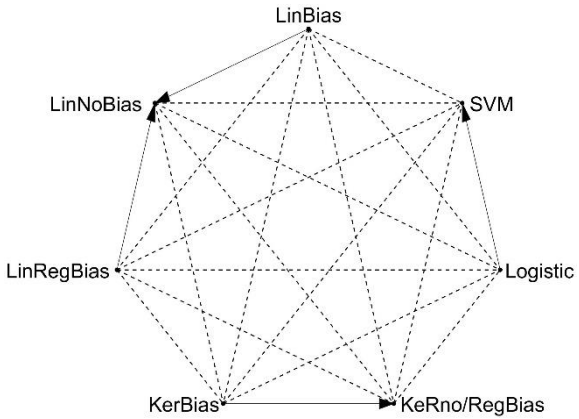
row vs. col.	$\mathcal{A}_1$	$\mathcal{A}_2$	$\mathcal{A}_3$	$\mathcal{A}_4$
$\mathcal{A}_1$	-	win	loss	tie
$\mathcal{A}_2$	loss	-	loss	tie
$\mathcal{A}_3$	win	win	-	win
$\mathcal{A}_4$	tie	tie	loss	-

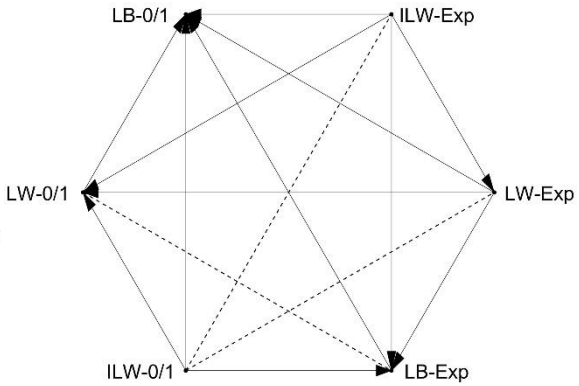
Left: The minimum N that makes the value of critical difference (CD) be less than 1 in Nemenyi post-hoc test, where *N*, *k* and *α* denote the number of datasets, the number of algorithms and the significance level, respectively; Right: An example using a table to show pairwise statistical comparative results of non-average rank-based strategies which is not very clear.

Main Contributions

- Contributions:
  - Analyze that the value of critical difference (CD) is often greater than one in our daily machine learning research, then the combination of the Friedman test and the Nemenyi post-hoc test usually leads to test results that are inconsistent with common sense;
  - Design a new presentation plot based on equilateral polygon to present win/tie/loss results. Specifically, each vertex corresponds to one algorithm. If one algorithm achieves statistically better performance than another, an arrow is used to connect them from the better one to the other. Otherwise, a dashed line connects them.







The proposed presentation plot. Left: Example in previous slide; Middle: Table 3(a) in [Doi: 10.1007/S13042-024-02114-6]; Right: Table IV (Accuracy) in [Doi:10.1109/TNNLS.2024.3454598].