Data-Driven Preference-Based Deep Statistical Ranking for Comparing Multi-objective Optimization Algorithms

Tome Eftimov[1,2(✉)], Peter Korošec[1,3], and Barbara Koroušić Seljak[1]

[1] Computer Systems Department, Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia
{tome.eftimov,peter.korosec,barbara.korousic}@ijs.si
[2] Jožef Stefan Postgraduate School, Jamova cesta 39, 1000 Ljubljana, Slovenia
[3] Faculty of Mathematics, Natural Sciences and Information Technologies, University of Primorska, Glagoljaška ulica 8, 6000 Koper, Slovenia

Abstract. To find the strengths and weaknesses of a new multi-objective optimization algorithm, we need to compare its performance with the performances of the state-of-the-art algorithms. Such a comparison involves a selection of a performance metric, a set of benchmark problems, and a statistical test to ensure that the results are statistical significant. There are also studies in which instead of using one performance metric, a comparison is made using a set of performance metrics. All these studies assume that all involved performance metrics are equal. In this paper, we introduce a data-driven preference-based approach that is a combination of multiple criteria decision analysis with deep statistical rankings. The approach ranks the algorithms for each benchmark problem using the preference (the influence) of each performance metric that is estimated using its entropy. Experimental results show that this approach achieved similar rankings to a previously proposed method, which is based on the idea of the majority vote, where all performance metrics are assumed equal. However, as it will be shown, this approach can give different rankings because it is based not only on the idea of counting wins, but also includes information about the influence of each performance metric.

Keywords: Multiple criteria decision analysis
Multi-objective optimization · Quality indicators
Deep statistical ranking · Statistical comparison · Data-driven

1 Introduction

When working on a new optimization algorithm, a crucial task is to compare its performance with state-of-the-art algorithms [1]. In single-objective optimization, the performance of algorithms is analyzed using the best algorithmic solution. For example, in the case of minimization problems, the solution with the

P. Korošec et al. (Eds.): BIOMA 2018, LNCS 10835, pp. 138–150, 2018.
https://doi.org/10.1007/978-3-319-91641-5_12

lowest value is the best. However, in multi-objective optimization algorithms (MOAs), it is not clear what the quality of a solution means in the presence of several optimization criteria. This is because the result is an approximation of the *Pareto-optimal* front, called an approximation set, which can be analyzed according to different quality aspects related to properties of convergence and diversity e.g., the closeness to the optimal front, coverage of a wide range of diverse solutions [2]. Quality indicators can be used to evaluate the performance of MOAs. Each quality indicator maps an approximation set to a real number [3]. In comparative studies, algorithms are used to solve a number of benchmark problems followed by the application of quality indicators to assess their performance [1]. Meta-heuristics are non-deterministic techniques, meaning there is no guarantee that the result will be the same for every run. To test the quality of an algorithm, it is not enough to perform just one run, but many runs of the algorithm on the same problem are needed, from which conclusions can be drawn. Additionally, this data must be analyzed with some statistical tests to ensure that the results are significant.

The aim of this study is to compare the performance of MOAs using a data-driven preference-based approach with a set of quality indicators. In Sect. 2, an overview of the related works is presented. Section 3 introduces the data-driven preference-based methodology. In Sect. 4 the experimental study is presented, while Sect. 4.3 gives a discussion of the proposed methodology. The conclusions of the paper are presented in Sect. 5.

2 Related Work

Many studies that address the problem of how to compare approximation sets in a quantitative manner have been conducted. Riquelme et al. [3] presented a study of a large number of metrics for comparing the performance of different multi-objective optimization algorithms, and presented a review and an analysis of 54 multi-objective optimization metrics and a discussion about the advantages/disadvantages of the most cited metrics in order to give researchers sufficient information for choosing them. A lot of the presented metrics use quality indicators to evaluate the quality of the solutions. Additionally, after calculating the quality indicator of interest, the data must be analyzed using a statistical test to ensure that the results are significant [4,5]. In [6], Eftimov et al. presented a study on how to compare the performance of MOAs using quality indicators and a Deep Statistical Comparison (DSC) approach. They used the DSC approach because it gives more robust statistical results to compare MOAs regarding the data obtained for a single quality indicator. However, there are also studies that use more than one quality indicator to evaluate the performance of MOAs. In [7], Yen and He presented a double-elimination tournament using a quality indicator ensemble to rank MOAs. The tournament contains approximation sets obtained from MOAs for the same initial population and involves a series of binary tournament selections and in each one a quality indicator from an ensemble is randomly chosen for comparison. The result of the tournament is one winning approximation set, so the corresponding MOA is ranked one. Then the approximations sets

that are generated by the winning MOA are removed and the remaining approximation sets will go through another double elimination tournament to identify the second best algorithm and so on. The results of the evaluation show that the method is performing more or less as a majority vote. The same idea was used by Ravber et al. [8], where instead of double elimination tournament, they used the chess rating system based on the Glicko-2 system [9]. The comparison between two approximation sets was made by a randomly selected quality indicator from the ensemble. In both approaches, the selection of the quality indicator that is used for a binary tournament is random and comes from a uniform distribution, such that all quality indicators in the ensemble are equal. Eftimov et al., also presented a comparative study of MOAs using an ensemble of quality indicators together with DSC [10]. This study used two ensemble combiners to rank and compare MOAs. Using one of them, each algorithm obtains a ranking for each problem, which is the average of its DSC rankings for each quality indicator for that problem. The other proposed ensemble is a hierarchical majority vote, which is a recursive approach where each algorithm is checked for the number of wins. In both scenarios, there is no preference between the quality indicators used in the comparison and all are assumed equal.

2.1 The Deep Statistical Ranking

Deep Statistical Comparison (*DSC*) is a recently proposed approach for making a statistical comparison of meta-heuristic stochastic optimization algorithms on a set of single-objective problems [4]. Its main contribution is its ranking scheme, which is based on the whole distribution instead of using just one statistic to describe the distribution, such as either the average or the median. A study on how to compare the performance of MOAs using quality indicators and DSC can be found in [6,10], where DSC gave more robust results compared to a standard statistical test recommended for making a statistical comparison.

2.2 The PROMETHEE

PROMETHEE methods are used in decision making to solve a decision problem in which a set of alternatives are evaluated according to a set of criteria that are often conflicting. Without loss of generality, we can assume that these criteria have to be minimized. For the method, an evaluation matrix is constructed, in which each alternative is estimated for each criteria. The method performs pairwise comparisons between all the alternatives for each criteria to provide either a complete or partial rankings of the alternatives. Four PROMETHEE methods exist, named as I, II, III, and IV. They can be used depending on the nature of the data that is involved in the comparison and the type of ranking that is preferred.

3 The Proposed Methodology

The proposed methodology consists of two steps. In the first, the DSC ranking scheme is used to obtain robust statistics regarding each quality indicator

separately, which are combined in the second step using the PROMETHEE II method [11].

3.1 The PROMETHEE II

Let us assume that a comparison needs to be made between m algorithms (i.e., alternatives) regarding n quality indicators (i.e., criteria) for a single problem. Let $A = \{A_1, A_2, \ldots, A_m\}$ be the set of algorithms we want to compare regarding the set of quality indicators $Q = \{q_1, q_2, \ldots, q_n\}$. The decision matrix is a $m \times n$ matrix (see Table 1) that contains the DSC rankings obtained for the algorithms for each quality indicator separately.

Table 1. Decision matrix.

	q_1	q_2	$\ldots$	q_n
A_1	$q_1(A_1)$	$q_2(A_1)$	$\ldots$	$q_n(A_1)$
A_2	$q_1(A_2)$	$q_2(A_2)$	$\ldots$	$q_n(A_2)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	
A_m	$q_1(A_m)$	$q_2(A_m)$	$\ldots$	$q_n(A_m)$

The DSC ranking scheme always ranks the best algorithm as one, the second best as two, and so on. In our case, we are interested in minimizing the criteria since lower DSC ranking values are preferable. Before we start with the PROMETHEE, the decision matrix is transformed in such a way that the DSC rankings, which are in the same column, are transformed using a standard competition ranking scheme [10]. This should be done because for the DSC rankings it does not matter if rankings are 1.50, 3.00, and 1.50 or 1.00, 3.00, and 1.00. In both scenarios having 1.00 and 1.50 means that the algorithm is the best according to some quality indicator. Since the DSC ranking scheme can never give a 1.00, 3.00, and 1.00 when comparing three algorithms (since it follows the idea of fractional ranking), the DSC rankings for each quality indicator are transformed using the standard competition ranking scheme.

The appropriate method in our case is PROMETHEE II. It is based on pairwise comparisons that need to be made between all algorithms for each quality indicator. The differences between DSC rankings for each pair of algorithms according to a specified quality indicator are taken into consideration. For larger differences the decision maker might consider larger preferences. The preference function of a quality indicator for two algorithms is defined as the degree of preference of algorithm A_1 over algorithm A_2 as seen in the following equation:

$$P_j(A_1, A_2) = \begin{cases} p_j(d_j(A_1, A_2)), & \textit{if maximizing the quality indicator} \\ p_j(-d_j(A_1, A_2)), & \textit{if minimizing the quality indicator} \end{cases}, \quad (1)$$

where $d_j(A_1, A_2) = q_j(A_1) - q_j(A_2)$ is the difference between the DSC rankings of the algorithms for the quality indicator q_j and $p_j(\cdot)$ is a generalized preference function assigned to the quality indicator. There exist six types of generalized preference functions [11]. In our case, usual preference function is used for each quality indicator because of the importance of any differences between the rankings, which is presented in Eq. 2.

$$p(x) = \begin{cases} 0, & x \leq 0 \\ 1, & x > 0 \end{cases}, \tag{2}$$

After selecting the preference function for each quality indicator, the next step is to define the average preference index and outranking (preference and net) flows. The average preference index for each pair of algorithms gives information of global comparison between them using all quality indicators. The average preference index can be calculated as:

$$\pi(A_1, A_2) = \frac{1}{n} \sum_{j=1}^{n} w_j P_j(A_1, A_2), \tag{3}$$

where w_j represents the relative significance (weight) of the j^{th} quality indicator. The higher the weight value of a given quality indicator the higher its relative significance. The selection of the weights is a crucial step in the PROMETHEE II method because it defines the priorities used by the decision-maker. In our case, we used the Shannon entropy weight method, which will be explained in the next subsection. For the average preference index, we need to point out that it is not a symmetric function, so $\pi(A_1, A_2) \neq \pi(A_2, A_1)$.

To rank the algorithms, the net flow for each algorithm needs to be calculated. It is the difference between the positive preference flow, $\phi(A_i^+)$, and the negative preference flow of the algorithm, $\phi(A_i^-)$. The positive preference flow gives information how a given algorithm is globally better than the other algorithms, while the negative preference flow gives the information about how a given algorithm is outranked by all the other algorithms. The positive and the negative preference flows are defined as:

$$\begin{aligned} \phi(A_i^+) &= \frac{1}{(n-1)} \sum_{x \in A} \pi(A_i, x), \\ \phi(A_i^-) &= \frac{1}{(n-1)} \sum_{x \in A} \pi(x, A_i). \end{aligned} \tag{4}$$

The net flow of an algorithm is defined as:

$$\phi(A_i) = \phi(A_i^+) - \phi(A_i^-). \tag{5}$$

The PROMETHEE II method ranks the algorithms by ordering them according to decreasing values of net flows.

3.2 The Shannon Entropy Weighted Method

To find the quality indicator weights, we use the Shannon entropy weighted method [12]. For this reason, the decision matrix presented in Table 1 needs to be normalized. Because the smaller value is preferred, the matrix is normalized using the following equation:

$$q_j(A_i)' = \frac{\max_i(q_j(A_i)) - q_j(A_i)}{\max_i(q_j(A_i)) - \min_i(q_j(A_i))}, \tag{6}$$

where $q_j(A_i)'$ is the normalized value for $q_j(A_i)$.

The entropy for each quality indicator is defined as:

$$e_j = K \sum_{i=1}^{m} W\left(\frac{q_j(A_i)'}{D_j}\right), \tag{7}$$

where D_j is the sum of the j^{th} quality indicator in all algorithms, $D_j = \sum_{i=1}^{m} q_j(A_i)'$, K is the normalized coefficient, $K = \frac{1}{(e^{0.5}-1)m}$, and W is a function defined as $W(x) = xe^{(1-x)} + (1-x)e^x - 1$.

The weight of each quality indicator used in Eq. 3 is calculated using the following equation:

$$w_j = \frac{\frac{1}{(n-E)}(1-e_j)}{\sum_{j=1}^{n}\left[\frac{1}{(n-E)}(1-e_j)\right]}, \tag{8}$$

where E is the sum of entropies, $E = \sum_{j=1}^{n} e_j$.

4 Results

4.1 Experimental Setup

The data from six algorithms is available from [13]. The algorithms are compared using 16 test problems. The number of objectives is set to four. More about the parameters of the test problems and the algorithms can be found in [13]. All test problems assume minimization of all objectives. Each algorithm was run for each problem 30 times. Before calculating the quality indicators, each approximated *Pareto* front was normalized. In our experiment quality indicators are hypervolume (q_1), epsilon indicator (q_2), r_2 indicator (q_3), and generational distance (q_4). All of them are unary indicators. Since we are introducing a methodology, we are not specifically dealing which quality indicators are used. The selection is up to user to make sure that relevant quality indicators are selected (e.g., if all quality indicators should be Pareto compliant, convergence, diversity, etc.). For calculating the hypervolume, the reference point $(1, \ldots, 1)$ is used, while for the other quality indicators, the reference set consists of all non-dominated solutions already known from all runs for each algorithm for a given problem. Because the DSC ranking scheme involves a statistical test for comparing distributions, a two-sample *Anderson-Darling (AD)* test is used and the significance level is set to 0.05. The benefits of using this test are presented in [14].

4.2 Experimental Results

In the experiment, three out of six algorithms are randomly selected. The algorithms are: DEMO^{SP2}, DEMO^{NS-II}, and NSGA-II. First, for each quality indicator, the DSC ranking scheme is used to compare the quality indicator data for a single problem. Further, the DSC rankings obtained for each quality indicator and each problem are transformed using the standard competition ranking scheme (see Table 2). The highest ranked algorithm for each problem and each quality indicator has the best performance.

Table 2. Transformed DSC rankings for each quality indicator of the algorithms, $A_1 = \text{DEMO}^{SP2}$, $A_2 = \text{DEMO}^{NS-II}$, and $A_3 = \text{NSGA-II}$.

Problem	Hypervolume			r_2			Epsilon			Generational distance		
	A_1	A_2	A_3	A_1	A_2	A_3	A_1	A_2	A_3	A_1	A_2	A_3
DTLZ1	2.00	1.00	3.00	1.00	2.00	3.00	1.00	2.00	3.00	1.00	2.00	3.00
DTLZ2	2.00	1.00	3.00	3.00	1.00	2.00	2.00	1.00	3.00	2.00	1.00	3.00
DTLZ3	1.00	1.00	3.00	2.00	1.00	3.00	1.00	1.00	3.00	1.00	1.00	3.00
DTLZ4	1.00	2.00	3.00	1.00	2.00	2.00	1.00	2.00	3.00	1.00	2.00	3.00
DTLZ5	2.00	2.00	1.00	1.00	1.00	3.00	1.00	1.00	1.00	1.00	3.00	2.00
DTLZ6	2.00	1.00	3.00	2.00	1.00	3.00	2.00	1.00	3.00	1.00	2.00	3.00
DTLZ7	2.00	1.00	3.00	2.00	1.00	3.00	2.00	1.00	3.00	2.00	1.00	3.00
WFG1	1.00	2.00	3.00	1.00	2.00	3.00	1.00	2.00	3.00	1.00	3.00	2.00
WFG2	1.00	2.00	3.00	1.00	2.00	2.00	1.00	2.00	2.00	1.00	3.00	1.00
WFG3	1.00	3.00	2.00	1.00	2.00	2.00	1.00	2.00	2.00	1.00	2.00	2.00
WFG4	1.00	2.00	3.00	2.00	1.00	2.00	2.00	1.00	3.00	3.00	2.00	1.00
WFG5	3.00	2.00	1.00	3.00	1.00	1.00	1.00	3.00	2.00	3.00	2.00	1.00
WFG6	1.00	2.00	3.00	2.00	1.00	3.00	1.00	2.00	2.00	3.00	1.00	1.00
WFG7	1.00	2.00	3.00	2.00	1.00	3.00	1.00	2.00	2.00	3.00	2.00	1.00
WFG8	1.00	2.00	2.00	1.00	2.00	3.00	1.00	2.00	2.00	1.00	3.00	2.00
WFG9	1.00	2.00	2.00	1.00	1.00	3.00	1.00	2.00	2.00	3.00	2.00	1.00

Before we find the complete ranking of the algorithms, the weights of each quality indicator are calculated for each single problem using the Shannon entropy weighted method. The weights for all problems are presented in Table 3.

Then, the PROMETHEE II method is used to rank the algorithms for each problem. If the original decision matrix is involved in the PROMETHEE II calculations, the preference function that is used is the one for minimizing the quality indicator, while if the normalized matrix is used, the preference function is the one used to maximize the quality indicator. In our case, we have a set of three algorithms $A = \{A_1, A_2, A_3\}$ that need to be compared according to a set of four quality indicators $Q = \{q_1, q_2, q_3, q_4\}$. The rankings obtained for PROMETHEE II method are presented on the left side of Table 4. They are

Table 3. Weights for each quality indicator.

Problem	q_1	q_2	q_3	q_4	Problem	q_1	q_2	q_3	q_4
DTLZ1	0.25	0.25	0.25	0.25	WFG2	0.14	0.37	0.37	0.12
DTLZ2	0.25	0.25	0.25	0.25	WFG3	0.13	0.29	0.29	0.29
DTLZ3	0.24	0.28	0.24	0.24	WFG4	0.18	0.46	0.18	0.18
DTLZ4	0.18	0.46	0.18	0.18	WFG5	0.26	0.22	0.26	0.26
DTLZ5	0.57	0.20	0.00	0.23	WFG6	0.19	0.19	0.47	0.15
DTLZ6	0.25	0.25	0.25	0.25	WFG7	0.18	0.18	0.46	0.18
DTLZ7	0.25	0.25	0.25	0.25	WFG8	0.36	0.14	0.36	0.14
WFG1	0.25	0.25	0.25	0.25	WFG9	0.37	0.12	0.37	0.14

further compared with the rankings obtained by the average ensemble with the DSC rankings (DSC ensemble I) [10], presented in the middle part of Table 4 and the hierarchical majority vote with the DSC rankings (DSC ensemble II) [10], presented on the right side of Table 4. From it, we can see that the rankings obtained using PROMETHEE II with DSC differ from the rankings obtained using the average ensemble with DSC or the hierarchical majority vote with DSC only in two bolded problems: DTLZ5 and WFG7.

Table 4. Ensemble combiner for the algorithms: $A_1 = \text{DEMO}^{\text{SP2}}$, $A_2 = \text{DEMO}^{\text{NS-II}}$, and $A_3 = \text{NSGA-II}$.

Problem	PROMETHEE II			DSC ensemble I			DSC ensemble II		
	A_1	A_2	A_3	A_1	A_2	A_3	A_1	A_2	A_3
DTLZ1	1.00	2.00	3.00	1.00	2.00	3.00	1.00	2.00	3.00
DTLZ2	2.00	1.00	3.00	2.00	1.00	3.00	2.00	1.00	3.00
DTLZ3	2.00	1.00	3.00	2.00	1.00	3.00	2.00	1.00	3.00
DTLZ4	1.00	2.00	3.00	1.00	2.00	3.00	1.00	2.00	3.00
DTLZ5	**2.00**	**3.00**	**1.00**	**1.00**	**2.50**	**2.50**	**1.00**	**2.50**	**2.50**
DTLZ6	2.00	1.00	3.00	2.00	1.00	3.00	2.00	1.00	3.00
DTLZ7	2.00	1.00	3.00	2.00	1.00	3.00	2.00	1.00	3.00
WFG1	1.00	2.00	3.00	1.00	2.00	3.00	1.00	2.00	3.00
WFG2	1.00	3.00	2.00	1.00	3.00	2.00	1.00	3.00	2.00
WFG3	1.00	3.00	2.00	1.00	3.00	2.00	1.00	3.00	2.00
WFG4	2.00	1.00	3.00	2.00	1.00	3.00	2.00	1.00	3.00
WFG5	3.00	2.00	1.00	3.00	2.00	1.00	3.00	2.00	1.00
WFG6	1.00	2.00	3.00	2.00	1.00	3.00	2.00	1.00	3.00
WFG7	**1.00**	**2.00**	**3.00**	**1.50**	**1.50**	**3.00**	**1.00**	**2.00**	**3.00**
WFG8	1.00	2.50	2.50	1.00	2.50	2.50	1.00	2.50	2.50
WFG9	1.00	2.00	3.00	1.00	2.00	3.00	1.00	2.00	3.00

To see what happens on a single problem, let us focus on the DLTZ5 problem. The decision matrix and its normalization are presented at top of Table 5. The transformed DSC rankings for the r_2 indicator and the DLTZ5 problem are 1.00, 1.00, and 1.00. Further, there is a problem in the normalization process because the normalized rankings are indeterminate forms (i.e., 0/0) [15], so the weight or the relative significance of this quality indicator can not be calculated. However, according to this quality indicator and the obtained DSC rankings, the compared algorithms are the same and they are all winners. Let us suppose that the weight w_3 could be calculated in some way, then the part of the average preference index that is related to the q_3 indicator is a product of $w_3 P_3(A_{i_1}, A_{i_2})$, where $i_1, i_2 = 1, \ldots, m$ and $i_1 \neq i_2$. In this case, it will be zero and will not influence the average preference index, which is used for calculating the positive and negative flows. Because it can not provide any additional information, it is removed and the result will be the same as comparing the algorithms regarding the remaining quality indicators, which in our case are q_1, q_2, and q_4. By removing the r_3 indicator, the decision matrix and its normalization are presented at the bottom part of Table 5. The weights obtained using the Shannon entropy weighted method are 0.57, 0.20, and 0.23. The final rankings and the outranking flows are given on the left side od Table 6. On the right part of Table 6 the average preference indices that are used for calculating the positive and negative flows for DLTZ5 are presented.

Table 5. Decision matrices for DLTZ5.

Algorithm	Decision matrix				Normalized matrix			
	q_1	q_2	q_3	q_4	q_1	q_2	q_3	q_4
DEMO^{SP2}	2.00	1.00	1.00	1.00	0.00	1.00	0/0	1.00
DEMO^{NS-II}	2.00	1.00	1.00	3.00	0.00	1.00	0/0	0.00
NSGA-II	1.00	3.00	1.00	2.00	1.00	0.00	0/0	0.50
Algorithm	Decision matrix				Normalized matrix			
	q_1	q_2	q_3	q_4	q_1	q_2	q_3	q_4
DEMO^{SP2}	2.00	1.00	/	1.00	0.00	1.00	/	1.00
DEMO^{NS-II}	2.00	1.00	/	3.00	0.00	1.00	/	0.00
NSGA-II	1.00	3.00	/	2.00	1.00	0.00	/	0.50

Table 6. Outranking flows, PROMOTHEE II rankings, and average indices for DLTZ5.

Algorithm	ϕ^+	ϕ^-	ϕ	Ranking		$\pi(A_i, A_1)$	$\pi(A_i, A_2)$	$\pi(A_i, A_3)$
DEMO^{SP2}	0.11	0.10	0.01	2.00	$\pi(A_1, A_j)$	0.00	**0.08**	**0.14**
DEMO^{NS-II}	0.03	0.17	−0.14	3.00	$\pi(A_2, A_j)$	**0.00**	0.00	**0.06**
NSGA-II	0.23	0.10	0.13	1.00	$\pi(A_3, A_j)$	**0.19**	**0.27**	0.00

Using the decision matrix presented in Table 5, the rankings obtained using the average ensemble and the hierarchical majority vote are the same and are 1.00, 2.50, and 2.50. In the case of hierarchical majority vote, DEMO^{SP2} is ranked as first because it wins in three out of four quality indicators, while DEMO^{NS-II} and NSGA-II are ranked second (e.g., 2.5) because both are ranked first in the case of two quality indicators, then both are second in the case of one quality indicator and third in the case of one quality indicator. All quality indicators are assumed equal and the ranking is made by counting the number of wins. However, the obtained rankings using the data-driven preference-based approach are 2.00, 3.00, and 1.00, which are completely different from the other ensembles. From the left part of Table 6, we can see that NSGA-II has the highest positive flow. The question is why it is ranked first when DEMO^{SP2} has two wins. This happens because the quality indicators that are involved have a data-driven preference for each of them, which is obtained by the Shannon entropy weighted method. The quality indicators are ordered as q_1, q_4, q_2, (e.g, hypervolume, generational distance, and epsilon indicator), starting from the most significant one to the least significant one. The average preference indices between A_1 and A_3 that are used for calculating the positive and negative flows are:

$$\begin{aligned}\pi(A_1, A_3) &= \frac{1}{3}\left[0.57 \cdot 0 + \mathbf{0.20} \cdot \mathbf{1} + \mathbf{0.23} \cdot \mathbf{1}\right] = 0.14 \\ \pi(A_3, A_1) &= \frac{1}{3}\left[\mathbf{0.57} \cdot \mathbf{1} + 0.20 \cdot 0 + 0.23 \cdot 0\right] = 0.19\end{aligned} \tag{9}$$

Using the calculations presented in Eq. 9, we can see that the average preference index between NSGA-II and DEMO^{SP2} is 0.19 and it is a result of only one win regarding the quality indicator q_1, while the average preference index between DEMO^{SP2} and NSGA-II is 0.14 and it is smaller even though it is a result of two wins regarding q_2 and q_4. This happens because q_1 is the most significant and its weight is much more than the sum of the weights of q_2 and q_4. In our experiment, the proposed data-driven preference-based approach gives different rankings from the hierarchical majority vote only for DLTZ5. This happens because only on that problem the compared algorithms are the same regarding one of the used quality indicators, which is the r_3 indicator. However, if this happens for other single-problems, the rankings can also differ from the rankings obtained by a hierarchical majority vote.

Furthermore, the obtained rankings using PROMETHEE II with DSC can be used as input data for a multiple-problem scenario. The appropriate statistical test is the *Friedman test*. Using it, the obtained p-value is 0.00, so using a significance level 0.05, we can conclude that there is a statistical significant difference between the compared algorithms using a set of benchmark problems. When comparing MOAs, often more than three algorithms are involved in the comparison, or especially a new algorithm is compared with state-of-the-art algorithm as a multiple comparisons with a control algorithm. When the number of algorithms increases the DSC rankings can be affected when correcting the p-values to control the FWER. In such a scenario, it is better to use multiple *Wilcoxon tests*, one for each pairwise comparison and then combine the p-values

to find the actual p-value for the scenario. More about this scenario and the DSC approach is presented in [4]. If we are interested in to compare them using a data-driven preference-based approach, we just need to use PROMETHEE II with DSC instead of the original DSC ranking scheme to find the rankings for each pairwise comparison on each problem.

4.3 Discussion

Comparing the performance of a new MOA with the performance of state-of-the-art MOAs is a crucial task in order to find its strengths and weaknesses. Different performance metrics can be used for evaluation and they are usually combined with statistical tests to ensure that the results are significant. Several previously proposed approaches are focused on comparing MOAs using a set of quality indicators. They follow the idea of ensemble learning, but all of them assume that all quality indicators are equal. The performance metric and the way how the algorithms will be compared also depend on the user preference or the concrete application. For example, in our previous work, we presented an average ensemble and a hierarchical majority vote based on counting wins according to different quality indicators, but in this paper we proposed a data-driven preference-based approach that is a combination of PROMETHEE II and DSC ranking scheme. According to the user preference all involved quality indicators are still equal, but the data-driven preference changes this by using its entropy. Organizing the DSC rankings for each quality indicator and each problem into a decision matrix, the Shannon entropy weighted method is used to find the relative significance of each quality indicator for each problem. The relative significance of each quality indicator is related to its entropy, which is the amount of information conveyed by it. The experimental results have shown that the preference-based approach performs more or less as a hierarchical majority vote. However, it can give different rankings, and the algorithm can overrank another one even if it has a lower number of wins, but it wins in most preferred quality indicator(s). Also, if there is a quality indicator for which all compared algorithms perform the same (they all win), it does not have an influence in the comparison and it can be removed from the set of quality indicators. Comparing the hierarchical majority vote and data-driven preference-based ranking, we can say that the hierarchical majority vote is more appropriate in cases where the performance is estimated by counting wins and loses such as in the case of dynamic multi-objective optimization, otherwise data-driven preference-based ranking can be used in cases when the influence of each quality indicator is required.

5 Conclusion

In this paper, we presented a data-driven preference-based approach for comparing MOAs using a set of quality indicators. The approach is a combination of PROMETHEE II, which is a method in MCDA, and a DSC ranking scheme,

that gives more robust statistical results and is based on comparing distributions instead of using only one statistic to describe the data. We compared our method with previously proposed methods where all involved quality indicators are assumed equal. We have shown that our method performs similar to a hierarchical majority vote, but also can give different rankings regarding the influence of each quality indicator, which is its preference and is estimated according to its entropy.

Acknowledgments. This work was supported by the project from the Slovenian Research Agency (research core funding No. P2-0098) and from the European Union's Horizon 2020 research and innovation program under grant agreement No. 692286.

References

1. Durillo, J.J., Nebro, A.J., Alba, E.: The jMetal framework for multi-objective optimization: design and architecture. In: 2010 IEEE Congress on Evolutionary Computation (CEC), pp. 1–8. IEEE (2010)
2. Coello Coello, C.A., Lamont, G.B., Van Veldhuizen, D.A., et al.: Evolutionary Algorithms for Solving Multi-objective Problems, vol. 5. Springer, New York (2007). https://doi.org/10.1007/978-0-387-36797-2
3. Riquelme, N., Von Lücken, C., Baran, B.: Performance metrics in multi-objective optimization. In: Computing Conference (CLEI), 2015 Latin American, pp. 1–11. IEEE (2015)
4. Eftimov, T., Korošec, P., Seljak, B.K.: A novel approach to statistical comparison of meta-heuristic stochastic optimization algorithms using deep statistics. Inf. Sci. **417**, 186–215 (2017)
5. García, S., Molina, D., Lozano, M., Herrera, F.: A study on the use of non-parametric tests for analyzing the evolutionary algorithms behaviour: a case study on the CEC2005 special session on real parameter optimization. J. Heuristics **15**(6), 617–644 (2009)
6. Eftimov, T., Korošec, P., Koroušić Seljak, B.: Deep statistical comparison applied on quality indicators to compare multi-objective stochastic optimization algorithms. In: Nicosia, G., Pardalos, P., Giuffrida, G., Umeton, R. (eds.) MOD 2017. LNCS, vol. 10710, pp. 76–87. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-72926-8_7
7. Yen, G.G., He, Z.: Performance metric ensemble for multiobjective evolutionary algorithms. IEEE Trans. Evol. Compu. **18**(1), 131–144 (2014)
8. Ravber, M., Mernik, M., Črepinšek, M.: Ranking multi-objective evolutionary algorithms using a chess rating system with quality indicator ensemble. In: 2017 IEEE Congress on Evolutionary Computation (CEC), pp. 1503–1510. IEEE (2017)
9. Glickman, M.E.: Example of the Glicko-2 system. Boston University (2012)
10. Eftimov, T., Korošec, P., Seljak, B.K.: Comparing multi-objective optimization algorithms using an ensemle of quality indicators with deep statistical comparison approach. In: 2017 IEEE Symposium Series on Computational Intelligence (SSCI) Proceedings, pp. 2801–2809. IEEE (2017)
11. Brans, J.P., Vincke, P.: Note - a preference ranking organisation method: (the PROMETHEE method for multiple criteria decision-making). Manag. Sci. **31**(6), 647–656 (1985)

12. Boroushaki, S.: Entropy-based weights for multicriteria spatial decision-making. Yearb. Assoc. Pac. Coast Geogr. **79**, 168–187 (2017)
13. Tušar, T., Filipič, B.: Differential evolution versus genetic algorithms in multiobjective optimization. In: Obayashi, S., Deb, K., Poloni, C., Hiroyasu, T., Murata, T. (eds.) EMO 2007. LNCS, vol. 4403, pp. 257–271. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-70928-2_22
14. Eftimov, T., Korošec, P., Seljak, B.K.: The behaviour of deep statistical comparison approach for different criteria of comparing distributions. In: Proceedings of 9th International Joint Conference on Computational Intelligence. SCITEPRESS Digital Library (2017)
15. Gordon, S.P.: Visualizing and understanding l'hopital's rule. Int. J. Math. Educ. Sci. Technol. **48**(7), 1096–1105 (2017)