Comparing Multi-Objective Optimization Algorithms Using an Ensemble of Quality Indicators with Deep Statistical Comparison Approach

Tome Eftimov
Computer Systems Department
Jožef Stefan Institute
Jožef Stefan Postgraduate School
Jamova cesta 39
Ljubljana, Slovenia
Email: tome.eftimov@ijs.si

Peter Korošec
Computer Systems Department
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenia
Faculty of Mathematics, Natural Science and Information Technologies
Glagoljaška ulica 8
Koper, Slovenia
Email: peter.korosec@ijs.si

Barbara Koroušić Seljak
Computer Systems Department
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenia
Email: barbara.korousic@ijs.si

***Abstract*—This paper presents a study on making a statistical comparison of multi-objective optimization algorithms using an ensemble of quality indicators together with a deep statistical comparison (DSC) approach. The DSC approach has been recently proposed for statistically comparing meta-heuristic stochastic optimization algorithms for single-objective problems. The DSC ranking scheme is based on the whole distribution, rather than on one statistic such as either the average or the median. This study uses two ensemble combiners to rank and compare algorithms using the DSC ranking scheme for each quality indicator for a given problem. Experimental results performed using 3 multi-objective optimization algorithms on 16 test problems show that ensembles of quality indicators with transformed DSC rankings give more robust results than when the same ensembles are used with transformed rankings obtained by some standard ranking schemes.**

1. Introduction

In real-world applications, many problems involve multi-objective optimizaiton [1]. Multi-objective optimization is related to a mathematical optimization problem involving more than one objective function to be optimized simultaneously. Because no single solution exists that simultaneously optimizes each objective function, the objective functions are said to be conflicting and there exists a set of alternative solutions. Each solution that belongs to the set of alternative solutions is optimal in a manner that no other solution from the search space is superior to it when all objective functions are considered. These solutions are known as *Pareto-optimal* solutions and the set is the *Pareto-optimal* set. A representation of the *Pareto-optimal* set in the objective space is called the *Pareto-optimal* front. Multi-objective optimization algorithms (MOAs) [2] are assumed as powerful techniques for finding a good approximation to the *Pareto optimal* front. To date, many MOAs have been developed, but there is no guarantee that optimal tradeoffs will be identified. They can however, produce a good approximation i.e, a set of solutions that is close to the optimal front. Experimental analysis of the performance of any new algorithm is a crucial task, and its performance must be compared with state-of-the-art algorithms. The aim of the comparative studies is, therefore, to explain the strengths and weaknesses of certain approaches and to identify the most promising algorithms.

Working with single-objective optimization, the performance of meta-heuristic stochastic optimization algorithms is analyzed using the best algorithmic solution. For example, in the case of minimization problems, the solution with lowest value is the better one. However, in multi-objective evolutionary algorithms, it is not clear what the quality of a solution means in the presence of several optimization criteria. The obtained result from a MOA is usually an approximation of the *Pareto-optimal* front, called an approximation set, and is analyzed according to different quality aspects, for example the closeness to the optimal front, coverage of a wide range of diverse solutions. Quality is measured in terms of criteria that relate to properties of convergence and diversity.

A large number of quality indicators, so-called performance indicators or performance metrics, are used to compare the performance of multi-objective stochastic optimization algorithms in multi-objective optimization. These indicators map the approximation sets to a set of real numbers. The idea is to quantify quality differences between approximation sets by applying common metrics from mathematics that result in real numbers. Quality indicators can be either unary or binary, but in principle can take an arbitrary number of arguments. An unary indicator is a real number assigned to each approximation set that reflects a certain quality aspect. A binary indicator is a real number that is

978-1-5386-2726-6/17/$31.00 ©2017 IEEE

assigned to pairs of approximation sets. The drawback of using binary indicators is when more than two algorithms, m, $m > 2$, need to be compared using a single binary indicator. In such a case, $m(m-1)$ different indicator values are obtained, one per each pairwise comparison. This makes the interpretation of the results more difficult in comparison to the m values in the case of unary indicators.

In multi-objective optimization studies, researchers are often interested in comparing the performance of algorithms using a set of quality indicators. To consider the influence of each used quality indicator, ensemble learning can be applied [3]. In our case an ensemble learning involves generating and combining multiple quality indicators together. Different techniques for combing quality indicator results exist including for example, voting-based methods, regression-based methods, and simple statistics.

In comparative studies, algorithms are used to solve a number of benchmark problems followed by the application of quality indicators to assess thier performance [4]. Metaheuristics 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, therefore, it is not enough to perform just one run, but to perform many runs, from which conclusions can be drawn. By calculating quality indicators for each approximation set, multivariate data is transformed into univariate data. Additionally, this data must be analyzed with some statistical tests to ensure that the results are significant. If not, any conclusions drawn may be wrong because differences between the algorithms could have occurred by chance. Further, if algorithms need to be compared over multiple multi-objective problems, for each algorithm for each problem either the average or the median of the quality indicator data needs to be calculated. This value is then used as a representative value in a multiple-problem scenario. Unfortunately, this can have a negative affect on the results of a statistical test [5], because averaging is sensitive to outliers that need to be considered especially because the MOAs can have poor runs. Even when there are no poor runs, the averages can be in some ϵ-neighborhood, which is s set of all numbers whose distance from a number is less than some specified number ϵ, and the algorithms will obtain different rankings. Only in the case of ties, average rankings are assigned. To overcome this, medians are sometimes used because they are more robust to outliers. However, medians can also be in some ϵ-neighborhood, and accordingly the algorithms will obtain different rankings. It can happen that the distributions of the obtained quality indicator data from multiple runs on one problem for the compared algorithms are the same, the median values are in some ϵ-neighborhood, and because of this the algorithms will have the same ranking. It can also happen that the distributions are different, and the median values are in some ϵ-neighborhood, and because of this the algorithms need to obtain different rankings. If this is the case, then the rankings of the algorithms are obtained according to their averages or medians, so the algorithm which has lower or higher value for either the average or median, it depends from the quality indicator that is used, is the better one.

For these reasons, in our previous work [6], an approach was used that removes the sensitivity of the simple statistics to the data and enables the calculation of more robust statistics without fearing the influence of outliers or some errors inside the ϵ-neighborhood. This approach is known as *Deep Statistical Comparison (DSC)* approach and was developed to compare meta-heuristics stochastic optimization algorithms for single-objective problems. The term *deep statistics* derives from the ranking scheme that is based on the whole distribution instead of using some simple statistics such as either the averages or medians.

The aim of this study is an ensemble of quality indicators using DSC for multi-objective optimization. First, for each quality indicator, a DSC ranking scheme was used to compare the obtained quality indicator data between the algorithms for a single problem. By using DSC, each algorithm obtains its ranking, which is robust to outliers and some ϵ-neighbourhood. Then, by using an ensemble combiner of the obtained rankings by each quality indicator, the rankings of the algorithms on that problem are calculated. Further, the obtained rankings from the ensemble of quality indicators for a single problem can be used together with some statistical omnibus test for a multiple problem scenario. The rest of the paper is organized as follows. In Section II, an overview of the related works is presented. In Section III, the DSC approach is reintroduced. Section IV introduces the ensemble learning process that uses the obtain rankings by each quality indicator involved in the comparison. Section V presents the experimental study with the discussion of the obtained results. The conclusions of the paper are presented in Section VI.

2. Related work

There have been many studies addressing the problem of how to compare approximation sets in a quantitative manner. Some of them include unary indicators, while other studies are based on binary indicators [7]. Another approach has been to use an attainment function, which involves estimating the probability of attaining arbitrary goals in objective space from multiple approximation sets [8]. Riquelme et al. [7] 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 fifty-four multi-objective optimization metrics and a discussion about the advantages/disadvantages of the most citied metrics in order to give researchers sufficient information for choosing metrics is necessary. Additionally, after calculating the quality indicator of interest, the data must be analyzed using statistical tests [6], [9].

The idea of ensemble learning is used especially in the domain of machine learning algorithms [3]. However, it has also been used for optimization algorithms [10]. Such an ensemble usually consists of very different optimization algorithms, so it is able to solve more, different optimization problems compared to each of the optimization algorithms alone.

3. Deep Statistical Comparison

Deep Statistical Comparison (*DSC*) is a recently proposed approach for comparing of meta-heuristic stochastic optimization algorithms over multiple single-objective problems [6]. Its main contribution is its ranking scheme, which is based on the whole distribution, instead of using only one statistic to describe the distribution, such as either the average or the median. The approach consists of two steps. The first step uses a newly proposed ranking scheme to obtain data in order to make a statistical comparison. The ranking scheme is based on comparing distributions using a statistical test, such as, the two-sample *Kolmogorov-Smirnov test* or the two-sample *Anderson-Darling* test [11]. All pairwise comparisons between the compared algorithms must be made, and the obtained p-values are organized in a matrix. Further, because multiple pairwise comparisons are made, these p-values are corrected using the *Bonferroni correction* [9] in order to control the family-wise error, FWER [12]. The FWER is the probability of making one or more false discoveries, or type I errors, among all hypotheses when performing multiple hypotheses tests. The matrix is then checked for transitivity, and on this basis the algorithms obtain their rankings. In addition, some vectors that are involved in the DSC ranking scheme need to be ordered in ascending or descending order depending on the quality indicator that is used. In the case of a single-objective optimization, these vectors must be ordered in ascending order, because the lowest value is the best solution assuming a minimization problem. The second step is a standard omnibus statistical test, which uses data obtained by the *DSC* ranking scheme as the input data. By using the *DSC*, wrong conclusions resulting from the presence of outliers or misleading ranking scheme can be avoided.

4. Ensemble combiner

Ensemble learning is used to compare MOAs regarding a set of quality indicators. However, before defining ensembles, it is important to provide some details about the DSC ranking scheme. The DSC ranking scheme produces data to be used used by an omnibus statistical test, following the idea of fractional ranking used in statistical tests. In fractional ranking, items that compare equal receive the same ranking, which is the average of what they would have received under ordinal rankings. The benefit of using a DSC ranking scheme is that the obtained rankings are more robust because they are based on comparison of distributions. For example, if we compare three algorithms on a given problem and the obtained DSC rankings are 1.50, 3.00, and 1.50, it means that the first and the third algorithm perform equally and are better than the second algorithm. However, to compare algorithms regarding a set of quality indicators, we need to combine the DSC rankings obtained by each quality indicator for each algorithm for a given problem. An ensemble can work as a competition, so it can combine the results of each quality indicator for a given problem and at the end it will indicate which algorithm wins. When combining DSC rankings, it does not matter if the DSC rankings are 1.50, 3.00, and 1.50 or 1.00, 3.00, and 1.00 because having 1.00 or 1.50 means that the algorithm is the best regarding some quality indicator. Since DSC ranking scheme can never provide 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 must be transformed using a standard competition ranking scheme, which is the adopted ranking scheme from the literature used for competitions. In standard competition ranking, items that compare equally receive the same ranking with a gap is left in the rankings. The number of rankings that are left out in this gap is one less than the number of items that compared equally. Each item ranking is 1 plus the number of items ranked above it. Using the standard competition ranking, it means that when two (or more) competitors tie for a position in the ranking, the position of all those ranked below them is unaffected.

Let us suppose that m MOAs are involved in the comparison according to a set of quality indicators. Two ensemble combiners, as competitions between the algorithms with respect to a set of quality indicators, are proposed. Because both ensembles are defined as competitions, they use a transformed DSC rankings using a standard competition ranking scheme according to the following equation:

$$Rank_{\mathrm{T}} = Standard\ competition\ ranking(Rank), \quad (1)$$

where $Rank$ is a $1 \times m$ vector that contains the DSC rankings obtained for a given quality indicator for a given problem, and $Rank_{\mathrm{T}}$ is a $1 \times m$ vector that contains the transformed DSC rankings using the standard competition ranking.

The first ensemble combiner is based on the average of the transformed DSC rankings. Each algorithm obtains a ranking for each problem, which is the average of its transformed DSC rankings by each quality indicator for that problem. The algorithm with the lowest ranking is the best one. The disadvantage of this combiner is the sensitivity of the average to the presence of outliers and in order to avoid this the median can be used instead.

The second ensemble combiner is a hierarchical majority vote. First, the combiner checks which algorithm wins in the most quality indicators or which algorithm is ranked in the most number of times with the best transformed DSC ranking. If the winner is only one algorithm, it will be ranked number 1. Then the other algorithms are checked according to their transformed DSC rankings, starting the comparison again from the best transformed DSC ranking. If there exists more than one algorithm with the same number of wins, then these algorithms are checked according to which of them is better according to the next ranking, or which of them is, in the most cases, ranked with the next transformed DSC ranking. This is recursively repeated until all of the algorithms obtain their rankings.

5. Results and discussion

In this section, we start by explaining the experimental setup, followed by a presentation of two experiments. In the first experiment, the first ensemble combiner is used and in the second experiment, the second ensemble combiner is used.

5.1. Experimental setup

The experimental data is the same as that used in [13] in which data is available for six algorithms. For this study, three out of the six algorithms are randomly selected: DEMO^{SP2}, DEMO^{NS-II}, and NSGA-II. To compare the presented algorithms, 16 test problems are used. The first consists of the first seven DLTZ test problems in [14] and the second of the nine WFG test problems presented in [15]. The number of objectives is set to 4. More about the parameters of the test problems and the parameters of 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. The quality indicators included the hypervolume, epsilon indicator, r_2 indicator, and generational distance. All used quality indicators are unary indicators. For calculating the hypervolume, the reference point (1,. . . , 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, for our experiments, the two-sample *Anderson-Darling (AD)* test is used. The benefits of using it are presented in [11]. The significance level for the AD test is set to 0.05.

5.2. First experiment

In this experiment, an example of using an ensemble of quality indicators based on the average of the transformed DSC rankings is presented. First, for each quality indicator, the DSC ranking scheme is used to compare the obtained quality indicator data for a single problem. The obtained DSC rankings are presented in Table 1 and are further compared with the Friedman rankings. For Friedman rankings, for each algorithm an average of each quality indicator data is calculated over 30 runs for a problem and it is further used by the Friedman ranking scheme. The algorithm with the best average value obtains the ranking 1, the second best average the ranking 2, and so on. Only in the case of a tie, an average ranking will be assigned. The Friedman rankings are presented in Table 2.

Using tables 1 and 2, there are problems for which the obtained rankings by the DSC and the Friedman ranking schemes differ. For hypervolume the differences appear for the problems: DTLZ3, DTLZ5, WFG8, and WFG9, for r_2 indicator: DTLZ4, DTLZ5, WFG2, WFG3, WFG4, WFG5, and WFG9, for epsilon indicator: DTLZ3, DTLZ5, WFG2, WFG3, WFG6, WFG7, WFG8, and WFG9, and for generational distance: DTLZ3, WFG2, WFG3, and WFG6. This happens because in the case when the averages and the Friedman ranking scheme are used, the average of the quality indicator data for a given problem can be affected by the presence of outliers, or because it could be in some ϵ-neigbourhood along with the averages of the quality indicator data for the other algorithms for that problem. In this case, obtained rankings will be different (Table 2). Only in the case of ties will an average ranking be assigned. In the DSC approach, instead of comparing the averages, distributions of the quality indicator data obtained by the algorithms for a single problem are compared and if found to be the same, the algorithms will obtain the same ranking even in the case when their averages are in some ϵ-neighbourhood. This makes the DSC rankings more reliable, since they are more robust to outliers and some ϵ-neighbourhood. To show this difference, in Figure 1, cumulative distributions (step functions) of the hypervolume of 30 runs of each algorithm for the WFG8 problem and average values (horizontal lines) of the hypervolume of 30 runs of each algorithm for the same problem are presented. From Figure 1, we can see that no statistical significant

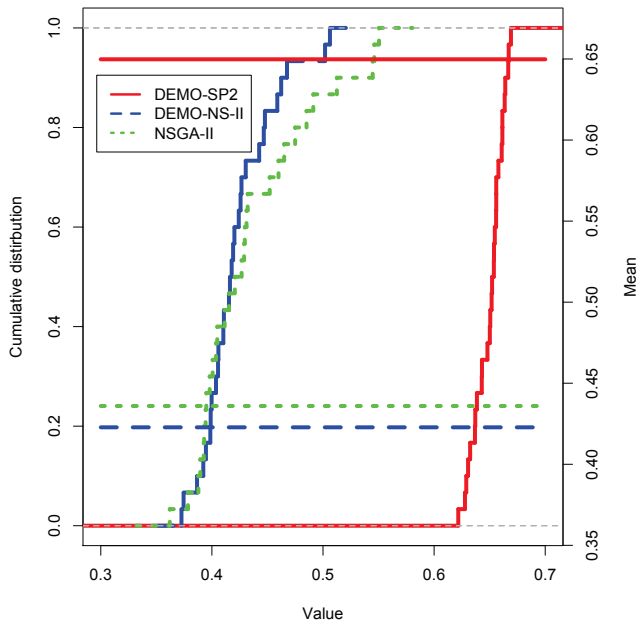

Figure 1. Cumulative distributions (step functions) and average values (horizontal lines) for hypervolume obtained for WFG8

difference exists between DEMO^{NS-II} and NSGA-II, so they receive the same ranking, but their cumulative distributions differ from the cumulative distribution of DEMO^{SP2}. Because a minimization problem of all four objectives is considered, the algorithm that has bigger hypervolume is the best, so it follows that DEMO^{SP2} will be ranked 1, and DEMO^{NS-II}, and NSGA-II will be ranked 2.50, despite the small difference that exists between their hypervolumes.

TABLE 1. DSC RANKINGS FOR EACH QUALITY INDICATOR OF THE ALGORITHMS, A_1=DEMO$^{\mathrm{SP2}}$, A_2=DEMO$^{\mathrm{NS-II}}$, AND A_3=NSGA-II

F	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	1.50	1.50	3.00	2.00	1.00	3.00	1.50	1.50	3.00	1.50	1.50	3.00
DTLZ4	1.00	2.00	3.00	1.00	2.50	2.50	1.00	2.00	3.00	1.00	2.00	3.00
DTLZ5	2.50	2.50	1.00	1.50	1.50	3.00	2.00	2.00	2.00	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	2.50	2.50	1.00	2.50	2.50	1.50	3.00	1.50
WFG3	1.00	3.00	2.00	1.00	2.50	2.50	1.00	2.50	2.50	1.00	2.50	2.50
WFG4	1.00	2.00	3.00	2.50	1.00	2.50	2.00	1.00	3.00	3.00	2.00	1.00
WFG5	3.00	2.00	1.00	3.00	1.50	1.50	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	2.50	2.50	3.00	1.50	1.50
WFG7	1.00	2.00	3.00	2.00	1.00	3.00	1.00	2.50	2.50	3.00	2.00	1.00
WFG8	1.00	2.50	2.50	1.00	2.00	3.00	1.00	2.50	2.50	1.00	3.00	2.00
WFG9	1.00	2.50	2.50	1.50	1.50	3.00	1.00	2.50	2.50	3.00	2.00	1.00

TABLE 2. FRIEDMAN RANKINGS FOR EACH QUALITY INDICATOR OF THE ALGORITHMS, A_1=DEMO$^{\mathrm{SP2}}$, A_2=DEMO$^{\mathrm{NS-II}}$, AND A_3=NSGA-II, OBTAINED ON AVERAGES OF EACH QUALITY INDICATOR OVER 30 RUNS

F	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	2.00	3.00	2.00	1.00	3.00	1.00	2.00	3.00	1.00	2.00	3.00
DTLZ4	1.00	2.00	3.00	1.00	2.00	3.00	1.00	2.00	3.00	1.00	2.00	3.00
DTLZ5	3.00	2.00	1.00	2.00	1.00	3.00	1.00	2.00	3.00	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	3.00	1.00	2.00	3.00	1.00	3.00	2.00
WFG3	1.00	3.00	2.00	1.00	3.00	2.00	1.00	2.00	3.00	1.00	3.00	2.00
WFG4	1.00	2.00	3.00	2.00	1.00	3.00	2.00	1.00	3.00	3.00	2.00	1.00
WFG5	3.00	2.00	1.00	3.00	1.00	2.00	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	3.00	3.00	2.00	1.00
WFG7	1.00	2.00	3.00	2.00	1.00	3.00	1.00	2.00	3.00	3.00	2.00	1.00
WFG8	1.00	3.00	2.00	1.00	2.00	3.00	1.00	3.00	2.00	1.00	3.00	2.00
WFG9	1.00	3.00	2.00	1.00	2.00	3.00	1.00	2.00	3.00	3.00	2.00	1.00

Using the averages values of the hypervolume on the same problem, the Friedman rankings will be 1.00 for DEMO$^{\mathrm{SP2}}$, 3.00 for DEMO$^{\mathrm{NS-II}}$ and 2.00 for NSGA-II.

Next, the first ensemble combiner is used, which is based on the average of the transformed rankings. This means that each algorithm will obtain a ranking for each problem, which is an average of its transformed rankings by each quality indicator for that problem. To see the differences, the ensemble combiner is used separately with a transformation of the DSC rankings and the Friedman rankings. Table 3 presents the transformed DSC rankings using the standard competition ranking scheme, while the transformed Friedman rankings using the standard competition ranking scheme remain the same (Table 2). The obtained rankings by the ensemble combiner are given on the left side of Table 4. The smallest ranked algorithm for each problem has the best performance. For example, for the problem WFG8 (Table 4) and the ensemble combiner based on the average of the transformed DSC rankings, the rankings of algorithms, DEMO$^{\mathrm{SP2}}$, DEMO$^{\mathrm{NS-II}}$, and NSGA-II, are 1.00, 2.250, and 2.250, respectively. This means, that for this problem, the performance of the DEMO$^{\mathrm{SP2}}$ differs from the performance of DEMO$^{\mathrm{NS-II}}$ and NSGA-II, so this algorithm has a better performance than the other two algorithms because it has smallest ranking, however there is no difference between the performance of the algorithms DEMO$^{\mathrm{NS-II}}$, and NSGA-II, because they have the same ranking according to the used ensemble of quality indicators. But this is not the case for transformed Friedman rankings where NSGA-II outperforms DEMO$^{\mathrm{NS-II}}$ according to Friedman rankings. For some problems, the rankings differ, but the orders are the same. There are, however, problems for which the orders also differ: DTLZ3, DTLZ5, WFG2, WFG6, WFG8, and WFG9. This happens because for those problems the rankings for some quality indicators by both ranking scheme differ, but the DSC ranking scheme gives more robust rankings.

For multiple problem scenario, where we compare the algorithms over multiple multi-objective problems, the obtained rankings by the ensemble combiner need to be used with an omnibus statistical test. This means, for making a comparison between the three algorithms, the Friedman test is the most appropriate because the required conditions for safe use of the parametric test are not satisfied. The test is used separately with the ensemble rankings obtained by each ranking scheme. In both cases, the p-value obtained by the Friedman test is 0.00, which means that there is a significant statistical difference between the performances of the algo-

TABLE 3. TRANSFORMED DSC RANKINGS FOR EACH QUALITY INDICATOR OF THE ALGORITHMS, A_1=DEMO$^{\mathrm{SP2}}$, A_2=DEMO$^{\mathrm{NS-II}}$, AND A_3=NSGA-II

F	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

TABLE 4. ENSEMBLE COMBINER FOR THE ALGORITHMS, A_1=DEMO$^{\mathrm{SP2}}$, A_2=DEMO$^{\mathrm{NS-II}}$, AND A_3=NSGA-II

F	Based on average						Hierarchical majority vote					
	DSC			Friedman			DSC			Friedman		
	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	1.250	1.750	3.000	1.250	1.750	3.000	1.000	2.000	3.000	1.000	2.000	3.000
DTLZ2	2.250	1.000	2.750	2.250	1.000	2.750	2.000	1.000	3.000	2.000	1.000	3.000
DTLZ3	1.250	1.000	3.000	1.250	1.750	3.000	2.000	1.000	3.000	1.000	2.000	3.000
DTLZ4	1.000	2.000	2.750	1.000	2.000	3.000	1.000	2.000	3.000	1.000	2.000	3.000
DTLZ5	1.250	1.750	1.750	1.750	2.000	2.250	1.000	2.500	2.500	1.000	2.000	3.000
DTLZ6	1.750	1.250	3.000	1.750	1.250	3.000	2.000	1.000	3.000	2.000	1.000	3.000
DTLZ7	2.000	1.000	3.000	2.000	1.000	3.000	2.000	1.000	3.000	2.000	1.000	3.000
WFG1	1.000	2.250	2.750	1.000	2.250	2.750	1.000	2.000	3.000	1.000	2.000	3.000
WFG2	1.000	2.250	2.000	1.000	2.250	2.750	1.000	3.000	2.000	1.000	2.000	3.000
WFG3	1.000	2.250	2.000	1.000	2.750	2.250	1.000	3.000	2.000	1.000	3.000	2.000
WFG4	2.000	1.500	2.250	2.000	1.500	2.500	2.000	1.000	3.000	2.000	1.000	3.000
WFG5	2.500	2.000	1.250	2.500	2.000	1.500	3.000	2.000	1.000	3.000	2.000	1.000
WFG6	1.750	1.500	2.250	1.750	1.750	2.500	2.000	1.000	3.000	1.000	2.000	3.000
WFG7	1.750	1.750	2.250	1.750	1.750	2.500	1.000	2.000	3.000	1.000	2.000	3.000
WFG8	1.000	2.250	2.250	1.000	2.750	2.250	1.000	2.500	2.500	1.000	3.000	2.000
WFG9	1.500	1.750	2.000	1.500	2.250	2.250	1.000	2.000	3.000	1.000	3.000	2.000

rithms according to the ensemble of quality indicators. At the single problem level, there are differences when using the ensemble with a transformation of the DSC rankings and the Friedman rankings. However, in this example, i.e, a multiple problem scenario, there is no difference in the results obtained. In general the differences at the single problem level can influence the result for a multiple problem scenario.

One weakness of this ensemble combiner is that the average of the obtained rankings for each quality indicator can be affected by the presence of outliers, especially when the number of compared algorithms increases. One way to avoid this is to use the median instead of the average because it is more robust to outliers. For this reason, another ensemble combiner is proposed, called the hierarchical majority vote.

5.3. Second experiment

In the second experiment, an ensemble combiner based on the hierarchical majority vote is presented. First, unique rankings from a set of all transformed rankings obtained from all quality indicators involved in the ensemble for a given problem are acquired. Then for each algorithm, it is counted how many instances of each of the selected unique rankings it has. The algorithm with the lowest ranking performs the best. Then, the combiner checks which algorithm is ranked in the most cases with the best ranking. If the winner is only one algorithm, it will obtain a ranking of 1, and the other algorithms need to be compared using their rankings starting again by comparing them from the best ranking. If there exist more than one algorithm that have the same number of wins, then these algorithms are compared with regards to the next ranking. In the case of same number of wins according to each unique ranking, the algorithms will obtain an average ranking. This is then recursively repeated until all algorithms obtain their rankings. The obtained rankings using the hierarchical majority vote with the transformed DSC and the transformed Friedman rankings are presented on the right side of Table 4. As in the first experiment, the algorithms rankings differ, when the ensemble combiner is used with the transformed DSC rankings and the transformed Friedman rankings. Looking at the right side of the Table 4, there are problems for which rankings on a single problem level differ: DTLZ3, DTLZ5,

TABLE 5. HIERARCHICAL MAJORITY VOTE FOR TWO PROBLEMS AND THE ALGORITHMS, A_1=DEMO$^{\mathrm{SP2}}$, A_2=DEMO$^{\mathrm{NS-II}}$, AND A_3=NSGA-II

DSC								Friedman							
DTLZ3				WFG8				DTLZ3				WFG8			
Ranking	A_1	A_2	A_3	Rank	A_1	A_2	A_3	Ranking	A_1	A_2	A_3	Rank	A_1	A_2	A_3
1.00	3	4	0	1.00	4	0	0	1.00	3	1	0	1.00	4	0	0
2.00	1	0	0	2.00	0	3	3	2.00	1	3	0	2.00	0	1	3
3.00	0	0	4	3.00	0	1	1	3.00	0	0	4	3.00	0	3	1
Final	2.00	1.00	3.00		1.00	2.50	2.50		1.00	2.00	3.00		1.00	3.00	2.00

WFG2, WFG6, WFG8, and WFG9. To see what happens for a single problem the hierarchical majority vote is presented in Table 5 for the problems, DTLZ3 and WFG8, for both ranking schemes, separately.

For the DTLZ3 problem, the unique rankings from the set of all transformed rankings using all quality indicators are selected for both ranking schemes. By using the transformed DSC rankings (Table 3), the unique rankings are 1.00, 2.00, and 3.00, and are the same for the transformed Friedman rankings. Then for each algorithm, the number of times each unique ranking is obtained is counted. Using the left side of Table 5, for the DTLZ3 problem, the algorithms are compared first according to which has the best ranking, which is 1. Accordingly, the DEMO$^{\mathrm{NS-II}}$ wins against the other two algorithms, i.e., it has the best transformed DSC ranking, and it is ranked 1. Then, the other two algorithms are compared. From this comparison the DEMO$^{\mathrm{SP2}}$ wins because it wins against the NSGA-II according to four quality indicators, so it obtains the final ranking 2, and the NSGA-II obtains the final ranking 3. The hierarchical majority vote for the same problem but with the transformed Friedman rankings is presented on the right side of Table 5. The unique rankings are 1.00, 2.00, and 3.00. In this case, the DEMO$^{\mathrm{SP2}}$ wins against the other two algorithms according to three quality indicators and it obtains a final ranking 1. The DEMO$^{\mathrm{NS-II}}$ is ranked 2nd because it wins against the algorithm NSGA-II according to the best transformed Friedman ranking. For the WFG8 problem, according to the best transformed DSC ranking, the DEMO$^{\mathrm{SP2}}$ wins against the other two algorithms according to four quality indicators and it obtains ranking 1. Then the other two algorithms are compared. Because they are the same according to the best transformed DSC ranking, the comparison continue with the next ranking. In this example, the other two algorithms are the same according to each unique ranking, so they obtain an average ranking, which means that there is no difference between their performance according to the set of used quality indicators. The hierarchical majority vote for the same problem with the transformed Friedman rankings is presented on the right side of Table 5.

For a multiple problem scenario, the obtained rankings by the ensemble combiner, called hierarchical majority vote, for a single problem must be used with an omnibus statistical test. The Friedman test, which is an appropriate omnibus statistical test, is used separately with the ensemble rankings obtained by each ranking scheme. In both cases, the obtained p-value by the Friedman test is 0.00, which means that there is a significant statistical difference between the performance of the algorithms according to the ensemble of the used quality indicators. The same as in the case of the first ensemble combiner at the single problem level, there are differences when using the ensemble with the transformed DSC rankings and the transformed Friedman rankings, while in a multiple problem scenario there is no difference of the obtained result. In general the differences that appear at the single problem level can influence the result for a multiple problem scenario.

In both experiment the examples involved three algorithms in the comparison. However, in a typical comparison often more than 3 algorithms are compared against the proposed one, as multiple comparisons with a control algorithm. In this case, when the number of algorithms increases, the DSC rankings can be affected by the correction of the p-values used in the DSC ranking scheme. To avoid this, more information about this scenario and the DSC can be found in [6].

6. Conclusion

To compare the performance of multi-objective evolutionary algorithms, a study for using an ensemble of quality indicators with deep statistical comparison (DSC) approach was presented. Two ensemble combiners have been proposed in order to rank and compare algorithms by using the rankings obtained by DSC for each quality indicator for a given problem. DSC is a novel approach for making a statistical comparison of meta-heuristic stochastic optimization algorithms over multiple single-objective problems. The main advantage of using DSC is that its ranking scheme, which instead of using only one statistic to describe the distribution, which can be average or median uses the whole distribution.

The evaluation of the study is performed by using the results for three multi-objective optimization algorithms tested on 16 test multi-objective problems. The obtained result are discussed on a single-problem analysis, when the algorithms are compared on a single multi-objective problem, and a multiple-problem analysis, when the algorithms are compared over multiple multi-objective problems. The study shows that ensembles of quality indicators with a transformation of the DSC rankings give more robust results than when the same ensembles are used with rank transformations by some standard ranking schemes. In a case of a single problem analysis, it can happen that there is a difference between the ensemble of quality indicators when the transformed DSC rankings and the transformed

rankings from some standard ranking scheme are used. By using the DSC ranking scheme, the obtained rankings are more robust to outliers or some ϵ-neighbourhood that exists between two numbers. In a multiple problem analysis, the experimental results show that there is no difference in the results according to which rankings are used. However, in general, the difference that exists on a single problem level can influence the end result of the multiple-problem analysis.

Acknowledgments

This work is supported by the project ISO-FOOD, which received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement No. 621329 (2014-2019). This work is part of a project that has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 692286. Also, we would like to thank Tea Tušar from the Department of Intelligent Systems at the Jožef Stefan Institute, for providing us with the experimental data, which is also public available on her website.

References

[1] K. Deb, K. Sindhya, and J. Hakanen, "Multi-objective optimization," in *Decision Sciences: Theory and Practice*. CRC Press, 2016, pp. 145–184.

[2] C. A. C. Coello, G. B. Lamont, D. A. Van Veldhuizen *et al.*, *Evolutionary algorithms for solving multi-objective problems*. Springer, 2007, vol. 5.

[3] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*. Springer, 2000, pp. 1–15.

[4] J. J. Durillo, A. J. Nebro, and E. Alba, "The jmetal framework for multi-objective optimization: Design and architecture," in *Evolutionary Computation (CEC), 2010 IEEE Congress on*. IEEE, 2010, pp. 1–8.

[5] T. Eftimov, P. Korošec, and B. Koroušić Seljak, "Disadvantages of statistical comparison of stochastic optimization algorithms," in *Proceedings of the Bioinspired Optimizaiton Methods and their Applications, BIOMA 2016*. JSI, 2016, pp. 105–118.

[6] ——, "A novel approach to statistical comparison of meta-heuristic stochastic optimization algorithms using deep statistics," *Information Sciences*, 2017.

[7] N. Riquelme, C. Von Lücken, and B. Baran, "Performance metrics in multi-objective optimization," in *Computing Conference (CLEI), 2015 Latin American*. IEEE, 2015, pp. 1–11.

[8] V. G. da Fonseca, C. M. Fonseca, and A. O. Hall, "Inferential performance assessment of stochastic optimisers and the attainment function," in *International Conference on Evolutionary Multi-Criterion Optimization*. Springer, 2001, pp. 213–225.

[9] S. García, D. Molina, M. Lozano, and F. Herrera, "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," *Journal of Heuristics*, vol. 15, no. 6, pp. 617–644, 2009.

[10] E. Yu and P. N. Suganthan, "Ensemble of niching algorithms," *Information Sciences*, vol. 180, no. 15, pp. 2815–2833, 2010.

[11] S. Engmann and D. Cousineau, "Comparing distributions: the two-sample anderson-darling test as an alternative to the kolmogorov-smirnoff test," *Journal of Applied Quantitative Methods*, vol. 6, no. 3, pp. 1–17, 2011.

[12] M. J. van der Laan, S. Dudoit, and K. S. Pollard, "Multiple testing. part ii. step-down procedures for control of the family-wise error rate," *Statistical applications in genetics and molecular biology*, vol. 3, no. 1, pp. 1–33, 2004.

[13] T. Tušar and B. Filipič, "Differential evolution versus genetic algorithms in multiobjective optimization," in *International Conference on Evolutionary Multi-Criterion Optimization*. Springer, 2007, pp. 257–271.

[14] K. Deb, L. Thiele, M. Laumanns, and E. Zitzler, *Scalable test problems for evolutionary multiobjective optimization*. Springer, 2005.

[15] S. Huband, L. Barone, L. While, and P. Hingston, "A scalable multi-objective test problem toolkit," in *International Conference on Evolutionary Multi-Criterion Optimization*. Springer, 2005, pp. 280–295.