Reducing Bias in Multi-Objective Optimization Benchmarking

Tome Eftimov
tome.eftimov@ijs.si
Computer Systems Department
Jožef Stefan Institute
Ljubljana, Slovenia

Peter Korošec
peter.korosec@ijs.si
Computer Systems Department
Jožef Stefan Institute
Ljubljana, Slovenia

ABSTRACT

The performance assessment of multi-objective optimization algorithms involves a user-preference-based selection of a single quality indicator used as a performance measure. A single quality indicator maps the approximation set (i.e., high-dimensional data) into a real value (i.e, one-dimensional data). However, it is well known that the selection of the quality indicator can have a huge impact on the benchmarking conclusions. This invites researchers to present only results for quality indicators that are in favor of the desired algorithm, or performing bias performance assessment. To go beyond this, we proposed a novel ranking scheme that reduces the bias in the user-preference selection by comparing the high-dimensional data of approximations sets and consequently provides more robust statistical results. The selection of a quality indicator is only required in cases when high-dimensional distributions of the approximation sets differ. By performing such analyses, experimental results show that the cases affected by the user-preference selection are reduced.

CCS CONCEPTS

• **Mathematics of computing → Hypothesis testing and confidence interval computation**.

KEYWORDS

multi-objective optimization, benchmarking, statistical analysis

ACM Reference Format:
Tome Eftimov and Peter Korošec. 2021. Reducing Bias in Multi-Objective Optimization Benchmarking. In *2021 Genetic and Evolutionary Computation Conference Companion (GECCO '21 Companion), July 10–14, 2021, Lille, France.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3449726.3462731

1 INTRODUCTION

Working with multi-objective optimization algorithms (MOAs) there is no one best solution, but an approximation of the Pareto front, which is called an approximation set. The approximation set is a set of high-dimensional solutions (i.e., the dimension comes from the number of objectives). The quality of the obtained set can be analyzed concerning different measures, known as quality indicators [5], concerning the convergence and diversity of the obtained solutions. The idea behind each quality indicator is that it transforms the high-dimensional data (i.e., the approximation set) into one-dimensional data (i.e., a real value). Since the quality indicators describe different aspects concerning the convergence and the diversity, there are studies that combined them as ensemble of quality indicators to provide more general conclusions [1, 3, 7].

Even though there are some insight into algorithms' performance by performing performance assessment using quality indicators, it is well known that the selection of a quality indicator(s) have impact to the end benchmarking conclusions [4]. Different quality indicators provide different explanations and results. Even more, not only the selection, but each quality indicator by transforming the high-dimensional into one-dimensional data losses information covered in the high-dimensional space that can affect the end result of the benchmarking. To go beyond this, we recently proposed a novel approach, known as Multi-objective Deep Statistical Comparison (moDSC) [2], which compares the distribution of the approximation sets (i.e., high-dimensional data). It reduces the information loss by transforming high-dimensional data into one-dimensional data. With this, it reduces the influence of the users' preference or the selection of a quality indicator to the end benchmarking conclusions. The quality indicator is applied only when the distributions of the high-dimensional data differ, indicating a statistically significant difference between approximation sets.

2 MULTI-OBJECTIVE DEEP STATISTICAL COMPARISON

MoDSC consists of two steps. In the first step, a ranking scheme that compares distributions of the solutions from the approximation sets (from different MOAs) is applied and ranks them for each benchmark problem involved in the study. While in the second step, the ranked data obtained for all benchmark problems is treated as input data for further analysis by an appropriate omnibus statistical test.

2.1 Ranking scheme

Let us assume that m MOAs should be compared using a set of n benchmark problems. Since these algorithms are stochastic in nature, there is no guaranty that the same approximation set will be obtained in each run, so each algorithm is run k times on each benchmark problem.

The moDSC ranking scheme involves $m(m-1)/2$ pairwise comparisons of the distributions of the approximation sets. These comparisons correspond to each pair of algorithms, where the distributions of the approximation sets of two MOAs obtained on the same benchmark problem are compared. To compare the distributions

GECCO '21 Companion, July 10–14, 2021, Lille, France

ACM ISBN 978-1-4503-8351-6/21/07.
https://doi.org/10.1145/3449726.3462731

within each benchmark problem, moDSC follows the idea of permutation tests, where M different pairs of approximation sets one per each MOA are compared with the multivariate ϵ test [6]. The M comparison are required since there is no one to one mapping between the approximation sets from both MOAs, each one has k different approximation sets obtained per each benchmark problem. Comparing M different pairs of approximations sets addresses the uncertainty presented in the data. This is supported by the assumption that all approximation sets obtained by a MOA on the same problem come from the same distribution.

By comparing M different pairs of approximation sets using the multivariate ϵ test, M p-values are obtained for each comparison of a pair of algorithms. To select one p-value from that comparison, a new random variable is introduced. This is a number of combinations for which the null hypothesis is rejected. To estimate if the compared distributions are either statistically significant or not, a significance level α_p must be selected. If $P(V) < \alpha_p$, both algorithms have the same distribution of the approximation sets, and vice-versa. If the distributions are not statistically significant, then a p-value is randomly selected from the subset of M p-values that are greater than α_X (i.e., significance level, in most cases 0.05). In opposite, the probability density function of a subset of M p-values that are lower than α_X is estimated using kernel density estimation. From it, its mode is selected as an appropriate p-value. In this case, the selection of the p-value is not random, since we are selecting from p-values lower than α_X, which can be further affected in the correction of p-values required to control the FWER (i.e., family-wise error rate).

By performing the above-mentioned procedure for each pairwise comparison (i.e., pair of algorithms) within the same benchmark problem, $m(m-1)/2$ p-values are obtained that can be organized into an $m \times m$ matrix. The rows and the columns correspond the the algorithms involved in the comparison. These values should be further corrected since multiple independent pairwise comparisons are performed. In our case, for illustration purposes, the Bonferroni correction is used, though more statistically powerful approaches can be also used. Further, this matrix can be transform to a binary matrix, where one corresponds that there is no statistical significance between the distributions of approximation sets of a pair of algorithms, and zero vice-versa. Checking the transitivity of the binary matrix will allow us to split it into disjoint sets of algorithms such that algorithms from the same set are not statistically significant concerning their distributions of the approximation sets. This leads that they should be ranked as the same and there is no reason to select a quality indicator. So, the quality indicator(s) should be selected only when the distributions differ. With this, the number of cases when the user preference (i.e., quality indicator(s)) should be used is reduced, which decreases the user-preference bias in the performance assessment.

2.2 Omnibus statistical test

Applying the ranking scheme for all n benchmark problems, the obtained rankings (an $n \times m$ matrix) can be further used and analyzed with an appropriate omnibus test to perform a multiple problem analysis and test the statistical significance between the performance of the MOAs.

3 DISCUSSION AND CONCLUSIONS

Most performance assessment studies that have been already conducted focus on comparisons done using one-dimensional data (i.e., quality indicators data). They all neglect the information presented in the high-dimensional space. Besides, many studies report quality indicator results that are tailored to the motivation in the development process of the algorithm, and as a consequence the algorithm wins. The problem that remains open is that all these performance assessment studies are bias to the user selection of which quality indicator will be applied.

The experimental results obtained by using the moDSC showed that it is able to reduce the cases when the user preference needs be selected. When the distributions of approximation sets are not statistically significant, they should be treated as the same and there is no reason to analyze them using quality indicator(s). The transformation from high-dimensional data to one-dimensional data can lead to completely different results. Reducing the number of cases that are affected by the selection of the user preference on a single-problem level leads to more robust statistical results when performing multiple-problem analysis. This comes from the fact that the rankings obtained for single benchmark problems affect the end test statistic of a multiple-problem analysis.

In future, we are planning to research random matrix theory approaches to estimate the rankings without involving the selection of quality indicators. This will be also supported by information theory approaches to estimate the amount/quantity of information that is lost during the transformation of high-dimensional data into one-dimensional.

ACKNOWLEDGMENTS

This work was supported by the financial support from the Slovenian Research Agency (research core funding No. P2-0098 and project No. Z2-1867).

REFERENCES

[1] Kalyanmoy Deb and Sachin Jain. 2002. Running performance metrics for evolutionary multi-objective optimization. (2002).

[2] Tome Eftimov and Peter Korošec. 2021. Deep Statistical Comparison for Multi-Objective Stochastic Optimization Algorithms. *Swarm and Evolutionary Computation* 61 (2021), 100837.

[3] Tome Eftimov, Peter Korošec, and Barbara Koroušić Seljak. 2017. 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*. IEEE, 2801–2809.

[4] Peter Korošec and Tome Eftimov. 2020. Multi-Objective Optimization Benchmarking Using DSCTool. *Mathematics* 8, 5 (2020), 839.

[5] Nery Riquelme, Christian Von Lücken, and Benjamin Baran. 2015. Performance metrics in multi-objective optimization. In *Computing Conference (CLEI), 2015 Latin American*. IEEE, 1–11.

[6] Gábor J Székely and Maria L Rizzo. 2004. Testing for equal distributions in high dimension. *InterStat* 5 (2004), 1–6.

[7] Heike Trautmann, Uwe Ligges, Jörn Mehnen, and Mike Preuss. 2008. A Convergence Criterion for Multiobjective Evolutionary Algorithms Based on Systematic Statistical Testing.. In *PPSN*. Springer, 825–836.