Robust Benchmarking for Multi-Objective Optimization

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

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

ABSTRACT

The performance assessment of multi-objective optimization algorithms is a crucial task for investigating their behaviour. However, the selected quality indicators and statistical techniques used in comparison studies can have huge impact on the study results. A quality indicator transforms high-dimensional data (an approximation set) into one-dimensional data (a quality indicator), followed by a potential loss of high-dimensional information concerning the transformation. Comparison approaches typically involve a single quality indicator or an ensemble of quality indicators to address more quality criteria, which are predefined by the user. To provide more robust benchmarking for multi-objective optimization, we extended the DSCTool with three approaches that are ensembles of quality indicators and one novel approach that compare the high-dimensional distributions of the approximation sets and reduces the users' preference in the selection of quality indicators. The approaches are provided as web services for robust ranking and hypothesis testing, including a proper selection of an omnibus statistical test and post-hoc tests if needed.

KEYWORDS

statistics, benchmarking, multi-objective optimization

ACM Reference Format:
Tome Eftimov and Peter Korošec. 2021. Robust Benchmarking for Multi-Objective Optimization. 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/10.1145/3449726.3463299

1 INTRODUCTION

Due to real-world relevance of multi-objective optimization, many multi-objective optimization algorithms have been developed. To analyze their behaviour, performance assessment is a crucial task. In contrast to the comparison of single-objective optimization algorithms, the multi-objective optimization algorithms do not obtain only one best solution for one problem, but a set of solutions called *approximation set*. Here, each solution from the approximation set is deemed optimal, so no other solution from the approximation set dominates it when all objectives are considered. For this purpose, different measures (in form of quality functions/indicators) have been proposed, which try to quantify the approximation sets concerning convergence and diversity. Each quality indicator maps an approximation set to a real number, making a transformation from high-dimensional data to one-dimensional data. Since quality indicators can describe only one characteristic of the approximation set, several approaches have been proposed that use more than one quality indicator (i.e., an *ensemble* of quality indicators to cover more quality aspects).

Since these algorithms are stochastic in nature, there is no guaranty that the same approximation set will be obtained in each run. For this purpose, they are run several times on the same problem and the obtained quality indicator data is analyzed with statistical tests. The analysis can be performed on a single problem or involving multiple problems.

In our recently published study, we have shown that the selection of the quality indicator can have huge impact on the results from the comparison study [8]. This allows users to select the quality indicator(s) in favour of their algorithm, leading to biased performance assessment. Even more, every transformation from high-dimensional data to one-dimensional data causes losing information from the high-dimensional space that could have influence on the comparison results.

To allow more robust benchmarking of multi-objective optimization algorithms, we have recently proposed three approaches of combining several quality indicators (i.e., ensembles of quality indicators), and a novel approach that reduces the potential bias in the selection of the quality indicators, known as moDSC. The three ensembles are: an average of quality indicators deep statistical rankings [4], a hierarchical majority vote [4], and a data-driven approach of fusing the quality indicators based on the information they conveyed (i.e., estimated by their entropy) [6].

For the purpose of Open Optimization Competition 2021, we have extended the DSCTool [7] in multi-objective optimization benchmarking scenarios, which consists of a set of REST web services that provide an understandable and error-free access to the power of Deep Statistical Comparison (DSC) [5]. We have also implemented R clients that can be used for easy integration with Nevergrad [1] and if possible with IOHprofiler [2] to support multi-objective benchmarking.

2 MULTI-OBJECTIVE DEEP STATISTICAL COMPARISON VARIANTS

Next, we provide a brief explanation of the three ensembles of quality indicators and the multi-objective DSC approach, which are included in the extension of the DSCTool.

The DSC ranking scheme is based on comparing one-dimensional quality indicator distributions using a two-sample statistical test with a predefined significance level.

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

ACM ISBN 978-1-4503-8351-6/21/07.
https://doi.org/10.1145/3449726.3463299

All ensembles of quality indicators involves users' selection of a set of quality indicators. Next, for each quality indicator separately, the DSC ranking scheme is used to rank the algorithms on each benchmark problem separately. Further, the obtained DSC rankings for each quality indicator are combined with three different heuristics and the obtained ranked data is further analyzed with an appropriate omnibus test.

2.1 Average ensemble of DSC rankings

The DSC average ensemble is a ranking scheme [4] that for a given pair of an algorithm and a benchmark problem calculates the average of the DSC rankings obtained for the set of quality indicators. This ensemble can be used to provide a more general conclusion in a benchmarking study.

2.2 Hierarchical majority vote based on DSC rankings

The DSC hierarchical majority vote ensemble is a ranking scheme [4] that counts which algorithm wins in the most quality indicators or which algorithm is ranked the most number of times with the best DSC ranking on each benchmark problem separately. This ensemble is recommended when dynamic multi-objective optimization is performed, where the performance can be treated as counting wins and loses.

2.3 Data-driven fusion using the DSC rankings

While the previous two ensembles treat the quality indicators with equal importance, the DSC data-driven ensemble [6] performs fusion of the quality indicators using the preference of each quality indicator estimated by its entropy. The weights (preferences) of the quality indicators are calculated by the Shannon entropy weighted method. The preference ranking organization method (PROMETHEE) is then used to determine the final rankings.

2.4 Multi-objective DSC

To reduce the potential bias in selecting the quality indicator(s), we have proposed a novel multi-objective Deep Statistical Comparison (moDSC) approach [3]. This approach directly compares the high-dimensional distributions of the approximation sets to determine if there is a statistical significance between them. Only in a case when a statistical significance is identified, a quality indicator is used to determine the appropriate ranking. The proposed approach is based on the Deep Statistical Comparison (DSC) [5] and it has been extended to analyse high-dimensional data. It consists of two steps. In the first step a novel multi-objective ranking scheme is used to rank the algorithms on each benchmark problem separately. The ranking is based on comparing high-dimensional distributions of approximation sets. In the second step, the ranked data is used as input data for an appropriate omnibus statistical test.

By using the moDSC, the effect of losing information while transforming high-dimensional to one-dimensional data is reduced. This indirectly reduces the user's preference in the quality indicator selection that can lead to bias comparison. Only when a statistical significance is observed in the distributions of the high-dimensional data, the user's preference is considered in the ranking process.

3 THE DEEP STATISTICAL TOOL

DSCTool offers web services that use a REST software architectural style that enables interoperability between different computer systems over the Internet. It consists of two steps: i) a mandatory one-time registration to access the DSCTool web services, and ii) selection of the desired ranking scheme (i.e., in our case *ensemble web services* or *multiobjective web service*). All ranking scheme web services provide data required to make a proper omnibus statistical test by calling the *omnibus web service*. If the null hypothesis is rejected, the web service provides data for proper selection and execution of a post-hoc test by calling the *posthoc web service*. A more detailed description of APIs can be accessed at https://ws.ijs.si:8443/dsc-1.5/documentation.pdf. The DSCTool is implemented in Java 1.8 and its web services are provided by Apache Tomcat 8.5.3 software. Source codes of the core library can be accessed at https://repo.ijs.si/korosec/dsc-core.git. For the Open Optimization Competition, we also provide R clients for easy integration with Nevergrad.

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] Pauline Bennet, Carola Doerr, Antoine Moreau, Jeremy Rapin, Fabien Teytaud, and Olivier Teytaud. 2021. Nevergrad: black-box optimization platform. *ACM SIGEVOlution* 14, 1 (2021), 8–15.
[2] Carola Doerr, Furong Ye, Naama Horesh, Hao Wang, Ofer M Shir, and Thomas Bäck. 2020. Benchmarking discrete optimization heuristics with IOHprofiler. *Applied Soft Computing* 88 (2020), 106027.
[3] Tome Eftimov and Peter Korošec. 2020. Deep Statistical Comparison for Multi-Objective Stochastic Optimization Algorithms. *Swarm and Evolutionary Computation* 61 (2020), 100837.
[4] Tome Eftimov, Peter Korošec, and Barbara Koroušić Seljak. 2017. Comparing multi-objective optimization algorithms using an ensemble of quality indicators with deep statistical comparison approach. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 1–8.
[5] Tome Eftimov, Peter Korošec, and Barbara Koroušić Seljak. 2017. A Novel Approach to statistical comparison of meta-heuristic stochastic optimization algorithms using deep statistics. *Information Sciences* 417 (2017), 186–215.
[6] Tome Eftimov, Peter Korošec, and Barbara Koroušić Seljak. 2018. Data-Driven Preference-Based Deep Statistical Ranking for Comparing Multi-objective Optimization Algorithms. In *International Conference on Bioinspired Methods and Their Applications*. Springer, 138–150.
[7] Tome Eftimov, Gašper Petelin, and Peter Korošec. 2020. DSCTool: A web-service-based framework for statistical comparison of stochastic optimization algorithms. *Applied Soft Computing* 87 (2020), 105977.
[8] Peter Korošec and Tome Eftimov. 2020. Multi-Objective Optimization Benchmarking Using DSCTool. *Mathematics* 8, 5 (2020), 839.