Optimization Letters
DOI 10.1007/s11590-006-0011-8

SHORT COMMUNICATION

How to assess and report the performance of a stochastic algorithm on a benchmark problem: *mean* or *best* result on a number of runs?

Mauro Birattari · Marco Dorigo

Received: 28 April 2006 / Accepted: 15 May 2006

Abstract Some authors claim that reporting the best result obtained by a stochastic algorithm in a number of runs is more meaningful than reporting some central statistic. In this short note, we analyze and refute the main argument brought in favor of this statement.

Keywords Assessment of performance · Experimental methodology · Stochastic algorithms · Metaheuristics

Notwithstanding the publication of a number of good methodological papers [3–6], many research works dealing with stochastic optimization algorithms still propose unsatisfactory empirical assessments. It is undeniable that empirical analyses play a major role in the study of stochastic optimization algorithms, in particular of metaheuristics for which gaining an analytical insight appears rather problematic. With this short note, we wish to address an apparently still open issue concerning how to summarize benchmark results.

In most research works, some index of performance of one or more stochastic algorithms on one or more benchmark problem instances is evaluated. When summarizing the results obtained by a given algorithm $\mathcal{A}$, rather that an indication of the central tendency of the observations, in unfortunately far too many cases, the *best* result b_N, observed in N runs, is reported. This quantity is not of any real interest. Indeed, it is just a particularly *over-optimistic* measure of

M. Birattari (✉) · M. Dorigo
IRIDIA-CoDE, Université Libre de Bruxelles, Brussels, Belgium
e-mail: mbiro@ulb.ac.be

M. Dorigo
e-mail: mdorigo@ulb.ac.be

the performance of the stochastic algorithm $\mathcal{A}$. Also, it should be noted that b_N is not a *good* estimator of the best performance that algorithm $\mathcal{A}$ can possibly achieve. In fact, it is well known that the empirical estimation of the maximum of a distribution (minimum if a minimization problem is considered) is particularly problematic. It is indeed always biased, since all possible observations are by definition smaller than or equal to the quantity to be estimated. Even worse, the uncertainty on the estimate does not nicely decrease with the size of the sample—as it does, for example, in the case of the estimation of the expected value.

Some authors, see for example Eiben and Jelasity [1] and Eiben and Smith [2], justify the use of b_N by saying that in a real-world application, if they were to find a good solution to the given problem instance, they would run their algorithm $\mathcal{A}$ for a number of times, say N, and then they would return the *best* solution found in these N runs. Although this seemingly reasonable argumentation contains some elements of truth, it is nevertheless faulty on a number of grounds. Some closer inspection is needed in order to disentangle the undeniable facts from the somehow dubious conclusions: on the one hand, it is perfectly legitimate to run $\mathcal{A}$ for N times and to use the best result found; in this sense, the authors of [1] and [2] are right when they maintain that a proper research methodology should take this widely adopted practice into account. On the other hand, the argumentation cannot be used for justifying the use of b_N as a measure of the performance of the algorithm $\mathcal{A}$. In the following, we accept the fact that one might wish to run $\mathcal{A}$ for N times for then selecting the best result and we highlight two main issues.

First of all, it should be recognized that, in this case, we are not discussing $\mathcal{A}$ but rather another algorithm, call it $\mathcal{A}^N$, which consists in *random restarting* $\mathcal{A}$ for N times. This entails two obligations. On the one hand, we should be clear from the beginning that we are interested in $\mathcal{A}^N$ and not in $\mathcal{A}$. On the other hand, if we are indeed interested in $\mathcal{A}^N$ we should provide a proper assessment of it. In particular, it should be recognized that by reporting the quantity b_N we are considering a *single* run of $\mathcal{A}^N$. Since $\mathcal{A}^N$ is itself a stochastic algorithm, an appropriate experimental methodology should study the central tendency of its performance: $\mathcal{A}^N$ should be observed over say M runs, which implies running NM times the underlying algorithm $\mathcal{A}$.

The second issue we wish to discuss is of a more subtle, if not provocative nature: the claim that we might be interested in $\mathcal{A}^N$ rather than in $\mathcal{A}$ should sound particularly suspicious when coming from researchers working on metaheuristics. A metaheuristic is typically understood [7–9] as a general-purpose method for guiding an underlying optimization algorithm, such as a problem-specific heuristic or a local search. In this sense, *random restart*, which consists in performing a number of say N independent runs of the underlying algorithm, can be seen as the most trivial metaheuristic: the *null*-metaheuristic. Indeed, many well-designed empirical analysis of metaheuristics include random restart as a performance yardstick: failing to improve over random restart is to be considered as a major failure for a metaheuristic. In the light of these considerations, it should sound strange that a researcher working in the metaheuristics field adopts a random restart strategy when he could have recourse to a more advance metaheuristic. This sounds somehow like betraying the fundamental principles of the research on metaheuristic.

Acknowledgments The authors acknowledge support from the ANTS project, an *Action de Recherche Concertée* funded by the Scientific Research Directorate of the French Community of Belgium. Marco Dorigo acknowledges support from the Belgian FNRS, of which he is a Research Director.

References

1. Eiben, A.E., Jelasity, M.: A critical note on experimental research methodology in EC. In: Proceedings of the 2002 Congress on Evolutionary Computation (CEC'2002), pp. 582–587. IEEE, New York (2002)
2. Eiben, A.E., Smith, J.E.: Introduction to Evolutionary Computing. Springer, Berlin Heidelberg New York (2003)
3. Fleming, P.J., Wallace, J.J.: How not to lie with statistics: the correct way to summarize benchmark results. Commun. ACM **29**(3), 218–221 (1986)
4. Barr, R.S., Golden, B.L., Kelly, J.P., Resende, M.G.C., Stewart, W.R.: Designing and reporting computational experiments with heuristic methods. J. Heuristics **1**(1), 9–32 (1995)
5. Hooker, J.N.: Testing heuristics: We have it all wrong. J. Heuristics **1**(1), 33–42 (1995)
6. Rardin, R.R., Uzsoy, R.: Experimental evaluation of heuristic optimization algorithms: a tutorial. J. Heuristics **7**(2), 261–304 (2001)
7. Stützle, T.G.: Local Search Algorithms for Combinatorial Problems – Analysis, Algorithms, and New Applications. PhD Thesis, Technische Universität Darmstadt (1999)
8. Glover, F., Kochenberger, G. (eds.): Handbook of Metaheuristics. Kluwer, Norwell (2002)
9. Hoos, H.H., Stützle, T.: Stochastic Local Search. Foundations and Applications. Morgan Kaufmann, San Francisco (2004)