Applied Soft Computing 19 (2014) 161–170

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Replication and comparison of computational experiments in applied evolutionary computing: Common pitfalls and guidelines to avoid them

Matej Črepinšek[a,*], Shih-Hsi Liu[b], Marjan Mernik[a]

[a] University of Maribor, Faculty of Electrical Engineering and Computer Science, Smetanova 17, 2000 Maribor, Slovenia
[b] California State University, Fresno, Department of Computer Science, 2576 E San Ramon Dr., Fresno, CA 93740, USA

ARTICLE INFO

Article history:
Received in revised form
13 December 2013
Available online 18 February 2014

Keywords:
Evolutionary algorithms
Experiments replication
Algorithms comparison
Economic load dispatch

ABSTRACT

Replicating and comparing computational experiments in applied evolutionary computing may sound like a trivial task. Unfortunately, it is not so. Namely, many papers do not document experimental settings in sufficient detail, and hence replication of experiments is almost impossible. Additionally, some work fails to satisfy the thumb rules for Experimentation throughout all disciplines, such that all experiments should be conducted and compared under the same or stricter conditions. Also, because of the stochastic properties inherent in evolutionary algorithms (EAs), experimental results should always be rich enough with respect to Statistics. Moreover, the comparisons conducted should be based on suitable performance measures and show the statistical significance of one approach over others. Otherwise, the derived conclusions may fail to have scientific merits. The primary objective of this paper is to offer some preliminary guidelines and reminders for assisting researchers to conduct any replications and comparisons of computational experiments when solving practical problems, by the use of EAs in the future. The common pitfalls are explained, that solve economic load dispatch problems using EAs from concrete examples found in some papers.

1. Introduction

"*If I have seen a little further, it is by standing on the shoulders of giants* [24]," this famous saying by Sir Isaac Newton, implies how important it is that theories should be established by correct and fair experimentation, so that pioneers may guide and assist subsequent researchers towards attaining even greater heights. In almost all science and engineering disciplines, the replication and comparison of experiments in a correct and fair manner is of utmost importance, in addition to the novelties and contributions (e.g., fast, accurate, robust, simple, high-impact, generalisable, and/or innovative [5]) regarding a newly proposed algorithm/method/methodology/technology/theory itself. Without such experimental results, researchers may be led along the wrong paths and evolution may be further delayed. Hence, the ability to build upon past results is crucial for progress in any science. It is interesting to point out that there has recently arisen a strong movement in computer science towards *computational reproducibility* [19,39]. The problem of reproducibility regarding experiments and the verifications of others' results in the field of EAs has also been identified by Eiben and Jelasity [17]. However, replications of computational experiments and good comparisons amongst different meta-heuristic methods [5] are difficult tasks. As noted by Barr et al. [5] "*When a new heuristic is presented in the computational and mathematical sciences literature, its contributions should be evaluated scientifically and reported in an objective manner, and yet this is not always done.*"

In order to tackle these issues, Barr et al. [5] presented discussions and comprehensive *general* guidelines as to how to design, compare, and report on computational experiments amongst different heuristic methods. Barr et al. also reminded researchers that reproducibility, specificity regarding all heuristic factors, preciseness of timing, availability of parameter settings, utilisation of Statistics, reduction in the variability of results, and the production of comprehensive results, are essential components needed in any report in order to assist future studies. After publishing Barr et al.'s work for more than 15 years, we are interested to know whether such issues have been addressed.

Despite more papers having been recently published on the designing and reporting on computational experiments (e.g., [6,17,36]) and on statistical methodology for comparing EAs (e.g.,

* Corresponding author. Tel.: +386 41736111.
E-mail addresses: matej.crepinsek@um.si (M. Črepinšek), shliu@CSUFresno.edu (S.-H. Liu), marjan.mernik@um.si (M. Mernik).

http://dx.doi.org/10.1016/j.asoc.2014.02.009
1568-4946/

[3,7,15,20,21,37]), we still come across numerous works where the replications and comparisons of computational experiments[1] are often improperly done, which can lead to improper conclusions whilst comparing different meta-heuristic methods. Moreover, if such experiments are then further used within other experiments, a rippling effect may occur and any comparison becomes worthless. It seems that the guidelines [5,6,17,36] have been mostly overlooked by practitioners in this field. One reason for overlooking these important works might be that practitioners often prefer concrete examples where the consequences of poorly replicating an experiment and unfair comparisons among algorithms could be clearly observed. Therefore, in this paper common pitfalls in experiment replication are discussed using concrete examples. In such a manner, common mistakes become more easily recognisable by practitioners and common pitfalls will be easier to avoid in the future. Note also that proper replication of experiments is a very fundamental and prerequisite for further comparisons of algorithms. If the experiment is not replicated with sufficient care, any performance measures and statistical approaches cannot remedy the problems introduced by inexact experiment replication. In other words, if collected data are gathered from experiments which exhibit large deviations the comparison is meaningless despite statistical test being applied. Hence, it is crucial that experiment replications are properly conducted. This paper is not about discussing which performance measures EA practitioners should use or which performance measure is the more appropriate. Although, in Section 2 a success rate (SR) [18] is used for particular examples. However, some performance measure (e.g., Mean Best Fitness (MBF) [18], average number of evaluations to a solution (AES) [18], expected running time (ERT) [23]) should be used during algorithm comparison in order to show the effectiveness of an approach.

This paper reviews a number of papers that have utilised EAs [18] to solve the economic load dispatch (ELD) problem [43], a practical real-world problem within those power plant operations whose objective functions try to allocate power generation to match load demand at minimal possible cost using specific system constraints. Our studies show that the pitfalls when replicating experiments and their comparisons are commonly seen in some of these papers. The main objective of this paper is not to criticise and challenge the results presented in these papers. Hence, all information which may point to particular works has been removed (e.g., authors' names, papers' titles, publishing information, algorithms' names). However, the removed data have been accessible to reviewers in the earlier versions of this paper enabling verification of data. More specifically, $Authors_1$ work was utilised as a benchmark for comparing their newly introduced algorithm for several papers. However, since the main purpose of $Authors_1$ work was to find only the best result (i.e., minimum cost) of the ELD problem, some important data for later experiment replications and comparisons were not shown in their paper. For example, (1) execution time was reported rather than the number of fitness evaluations. Such a measure, however, is subject to the hardware specifications of the platform conducting experiments; (2) minimum, mean, and maximum costs were reported, but the standard deviations for mean costs were missing, which makes experiment comparison less accurate with respect to Statistics; and (3) as for the performance measure a complex frequency of convergence was used rather than the simpler SR. The former measure is more difficult to interpret than the latter one and, therefore, was usually omitted in subsequent experiments. Although, there still can be some problems, as discussed later, with SR [11]. Additional drawbacks were also discovered in some of the subsequent works. For example, (1) only a partial test-suite was replicated and compared; (2) the number of independent runs were fewer than its benchmark paper; (3) large deviations in the number of fitness evaluations consumed; (4) whether the experiments utilised the best parameter settings was unreported; (5) not all performance measures were utilised; and (6) the statistical significance of the results was not shown. With such missing information and different experimental settings, experimental results and conclusions of the subsequent papers may become less convincing. The aforementioned problems are not only pertinent to the ELD problem, but also to many other studies solving practical problems (e.g., turning operations [38], milling operations [27], welded beam design [1], and pressure vessel design [1]). Hence, the ELD problem was chosen arbitrarily to designate the quite common problem of experiment replication and comparison in the field of applied evolutionary computing. On the other hand, we should point out that there are also numerous works (e.g., [2,25,26,32]) where performance measurements have been collected with variances, and statistical tests have been performed. Further recent competitions on real-parameter optimisations [23,28,40,41] provide an excellent experimental setup, which still needs to be accepted by EA practitioners. On the other hand, in all of these competitions, the algorithms were run on the same computer platform, and good performance measures were easier to define. But, the same computing environment is much harder to achieve by practitioners for practical industrial problems, and comparison is usually done based on reported results. It is also interesting to point out that many other applied sciences have problems performing empirical studies (e.g., Software Engineering [35], Medicine [44]). For example, Welch and Gabbe in [44] reported that more than half of the studies were undocumented regarding sufficient details, and replications were impossible. We chronologically detail the common pitfalls of experiment replications and comparisons found in ELD papers as a friendly reminder – through these studies, this paper extends comprehensive general guidelines from [5,6,17,36], and the guidelines from Črepinšek et al. [13] with special emphases on guidelines for replicating experiments in applied evolutionary computing. A checklist is also introduced to remind researchers to avoid such common pitfalls in the future.

This paper is organised as follows: Section 2 summarises the common pitfalls of EAs replications and comparisons on the ELD problem. The chronological development and drawbacks of the experiments are then further detailed. Section 3 offers the guidelines learned from Section 2, along with a checklist to assist researchers on the replication of experiments in applied evolutionary computing, followed by the conclusions in Section 4.

2. Case study: a practitioner's approach

EAs have been used, from [4,18,22,34]'s inception, for solving hard optimisation problems. Solving real-world problems is actually the ultimate purpose of any EAs. In this section, common pitfalls regarding experiment replication and comparison are explored and explained on the economic load dispatch (ELD) problem (also known as economic dispatch (ED) problem), which is inherently a high-nonlinear and non-convex problem [33,43,45]. The problem, allocating power generation to match load demand at minimal possible cost using specific system constraints, has been intensively studied for the last 20+ years (Scopus search on "economic load dispatch" performed on July 1, 2012, returns 1526 hits, whilst search on "economic dispatch" returns 3143 hits). This problem has recently been extended into a multi-criteria problem, also including a request for minimising emission levels [8].

Although the work by $Authors_1$ was not amongst the first for solving the ELD problem using EAs techniques, it has served as ground research for many other researchers who have used the

[1] In the continuation we will skip the term 'computational', but whenever 'experiment' is mentioned we will have 'computational experiment' in our minds.