ARTICLE IN PRESS

Information Sciences xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Information Sciences

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

A chess rating system for evolutionary algorithms: A new method for the comparison and ranking of evolutionary algorithms

Niki Veček*, Marjan Mernik, Matej Črepinšek

University of Maribor, Faculty of Electrical Engineering and Computer Science, Smetanova 17, 2000 Maribor, Slovenia

ARTICLE INFO

Article history:
Received 2 September 2013
Received in revised form 13 February 2014
Accepted 25 February 2014
Available online xxxx

Keywords:
Evolutionary algorithm
Computational experiment
Null hypothesis significance testing
Chess rating
Ranking

ABSTRACT

The Null Hypothesis Significance Testing (NHST) is of utmost importance for comparing evolutionary algorithms as the performance of one algorithm over another can be scientifically proven. However, NHST is often misused, improperly applied and misinterpreted. In order to avoid the pitfalls of NHST usage this paper proposes a new method, a Chess Rating System for Evolutionary Algorithms (CRS4EAs) for the comparison and ranking of evolutionary algorithms. A computational experiment in CRS4EAs is conducted in the form of a tournament where the evolutionary algorithms are treated as chess players and a comparison between the solutions of two algorithms on the objective function is treated as one game outcome. The rating system used in CRS4EAs was inspired by the Glicko-2 rating system, based on the Bradley–Terry model for dynamic pairwise comparisons, where each algorithm is represented by rating, rating deviation, a rating/confidence interval, and rating volatility. The CRS4EAs was empirically compared to NHST within a computational experiment conducted on 16 evolutionary algorithms and a benchmark suite of 20 numerical minimisation problems. The analysis of the results shows that the CRS4EAs is comparable with NHST but may also have many additional benefits. The computations in CRS4EAs are less complicated and sensitive than those in statistical significance tests, the method is less sensitive to outliers, reliable ratings can be obtained over a small number of runs, and the conservativity/liberality of CRS4EAs is easier to control.

1. Introduction

Computational experiments are essential within the field of meta-heuristics [8,9]. New meta-heuristic algorithms with different exploration and exploitation abilities [16] have been proposed and their effectiveness needs to be displayed on artificial or real problems by theoretical analysis or empirical testing. Empirical testing when performing computational experiments is preferred within meta-heuristics as theoretical analysis is based on mathematical theorems and models and can be more difficult to understand. Essentially, the contributions of a new meta-heuristic algorithm should be evaluated scientifically and reported objectively. As already noted by Barr et al. [5] this is not always the case for heuristic methods. We have also observed the following problems in some meta-heuristic literature: newly developed algorithms are not described in sufficient detail nor are publicly available, the experimental settings are only partially documented thus preventing an exact

* Corresponding author. Tel.: +386 40570780.
E-mail addresses: niki.vecek@uni-mb.si (N. Veček), marjan.mernik@uni-mb.si (M. Mernik), matej.crepinsek@uni-mb.si (M. Črepinšek).

http://dx.doi.org/10.1016/j.ins.2014.02.154
0020-0255/

replication of an experiment and fair comparison, experiments are not conducted under the same or stricter conditions as the original ones, the experimental results are not always rich enough with respect to statistics, improper use or even the absence of statistical methods, statistical significance is omitted, and the derived conclusions are too general and unsupported by experiments - these problems are also discussed in [25]. The aforementioned problems are not only typical for the field of meta-heuristics. For example, Kitchenham et al. [61] mentioned similar problems within the fields of software engineering and medicine, where it was reported that 40% of the examined medical publications (in total 164) had statistical errors; and even in another study half of the publications were impossible to evaluate due to insufficient details of the statistical methods used, whilst nearly one third of the publications contained inappropriate usages of statistics. If researchers have problems with the proper usages of statistical methods within an established discipline such as medicine, then it is unsurprising that researchers have problems in much younger disciplines such as software engineering [23] and meta-heuristics. This problem is vividly described in [61] as: *"Some problems with statistics arise because there are methodological difficulties applying standard statistical procedures to software experiments. Nonetheless, the majority of problems result from lack of statistical expertise in the empirical research community."*

As the main reason for the plethora of statistical errors is inadequate understanding of statistical methods, the question is how to alleviate this problem during computational experiments within the field of meta-heuristics, in particular evolutionary computations [4,26]. Better training might be an obvious answer but this solution would only come into effect long-term. In the meantime the proper designing, executing, and reporting of computational experiments will remain a crucial task. This paper describes a possible alternative solution in which knowledge of the Null Hypothesis Significance Testing (NHST) is not essential when comparing evolutionary algorithms. A novel Chess Rating System for Evolutionary Algorithms (CRS4EAs) method is proposed for the comparisons and rankings of evolutionary algorithms as a feasible alternative to NHST. Roughly speaking, the idea of rating chess players is incorporated within our method. Nowadays, chess ratings are very reliable regarding chess players' strengths [42,44]. Can chess ratings be used to measure the strengths of evolutionary algorithms, and hence used for algorithms' comparisons? This paper shows that CRS4EAs can indeed be used as an alternative for algorithm rankings and comparisons. The following analogies are used in CRS4EAs: 'chess player = evolutionary algorithm', 'chess game = searching for the best solution for a given problem using a pre-specified number of fitness evaluations', and 'chess tournament = comparison of evolutionary algorithms on the benchmark test suite using a pre-specified number of independent runs'. On a benchmark suite containing 20 standard numerical optimisation functions, 16 different evolutionary algorithms and their variants are compared using the newly-proposed CRS4EAs method, as well as classical NHST. It was shown that CRS4EAs has many similarities to NHST, whilst on the other hand it also has several additional benefits such as: (1) robustness to outliers, (2) a controllable mechanism for conservativity/liberality, (3) is simple to apply without the danger of misuse or misunderstanding, (4) has a simple experimental design and (5) accurate estimate of an algorithm's performance is achieved over a small number of runs (25 or less). The limitations of CRS4EAs can be seen in: (1) the number of total runs (i.e. games) depends on the number of algorithms compared, (2) the ranking list changes with each new added algorithm, (3) the means and other statistical values reported within existing publications cannot simply be re-used for algorithm comparisons, (4) a ranking-list with a slightly different experiment results must not be used identically for further comparing a new algorithm, or (5) the handling of the outliers might be less appropriate in situations that demand high success rates in all independent runs due to one-to-one run (game) comparison.

The main contributions of this paper are: the proposal of a new method CRS4EAs for comparing and ranking evolutionary algorithms, and the presentation of a computational experiment that empirically confirmed that CRS4EAs can be used for comparing evolutionary algorithms and is comparable with NHST but also contains several other benefits. We wish CRS4EAs to be used not only amongst researchers but to also assist reviewers in evaluating newly-developed algorithms.

This paper is organised as follows. Section 2 briefly reviews the classical method of evolutionary algorithms' comparisons by using various statistical methods. Section 3 presents an introduction to those chess rating systems currently in use by different Chess Federations with an emphasis on the Glicko and the Glicko-2 rating systems. The novel CRS4EAs method for ranking evolutionary algorithms is introduced in Section 4. Extensive experimental results after comparing the two approaches, CRS4EAs and NHST, for evolutionary algorithms' comparisons are presented in Section 5, followed by a discussion in Section 6. The paper concludes in Section 7 with a brief statement on proposed future directions.

2. Background

A common method for comparing the performances of different evolutionary algorithms is modern statistical hypothesis testing, which was developed by Fisher [33,34,36], and Neyman and Pearson [71]. After stating a null hypothesis that there is no difference in the results from the experiment, an appropriate statistical test is used to check whether the null hypothesis could be rejected or not. Any rejection of the null hypothesis would lead to acceptance of the alternative hypothesis that states that there would be differences in the results of the experiment. Hypothesis testing requires the specification of an acceptable level of statistical error. When the null hypothesis is falsely rejected a Type I error is made, and if the test fails to reject a false null hypothesis a Type II error is made. The probabilities of both errors are also known as α and β. The probability of correctly rejecting a false null hypothesis, i.e. $1-\beta$, is the power of statistical test.

Usually, the goal of statistical inference in evolutionary algorithms is to compare multiple algorithms on multiple data sets and there are several statistical techniques for doing this that are divided into two types – parametric and non-parametric

Please cite this article in press as: N. Veček et al., A chess rating system for evolutionary algorithms: A new method for the comparison and ranking of evolutionary algorithms, Inform. Sci. (2014), http://dx.doi.org/10.1016/j.ins.2014.02.154