Information Sciences 417 (2017) 186–215

Contents lists available at ScienceDirect

Information Sciences

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

A Novel Approach to statistical comparison of meta-heuristic stochastic optimization algorithms using deep statistics

Tome Eftimov[a,b,*], Peter Korošec[a,c], Barbara Koroušić Seljak[a]

[a] Computer Systems Department, Jožef Stefan Institute, Jamova cesta 39, Ljubljana 1000, Slovenia
[b] Jožef Stefan International Postgraduate School, Jamova cesta 39, 1000 Ljubljana, Slovenia
[c] Faculty of Mathematics, Natural Science and Information Technologies, Glagoljaška ulica 8, Koper 6000, Slovenia

ARTICLE INFO

Article history:
Received 5 October 2016
Revised 25 May 2017
Accepted 11 July 2017
Available online 14 July 2017

Keywords:
Statistical comparison
Stochastic optimization algorithms
Single objective function
Ranking scheme
Multiple problem analysis

ABSTRACT

In this paper a novel approach for making a statistical comparison of meta-heuristic stochastic optimization algorithms over multiple single-objective problems is introduced, where a new ranking scheme is proposed to obtain data for multiple problems. The main contribution of this approach is that the ranking scheme is based on the whole distribution, instead of using only one statistic to describe the distribution, such as average or median. Averages are sensitive to outliers (i.e., the poor runs of the stochastic optimization algorithms) and consequently medians are sometimes used. However, using the common approach with either averages or medians, the results can be affected by the ranking scheme that is used by some standard statistical tests. This happens when the differences between the averages or medians are in some ϵ-neighborhood and the algorithms obtain different ranks though they should be ranked equally given the small differences that exist between them. The experimental results obtained on Black-Box Benchmarking 2015, show that our approach gives more robust results compared to the common approach in cases when the results are affected by outliers or by a misleading ranking scheme.

1. Introduction

A comprehensive comparison of the efficiency of stochastic optimization algorithms has in recent years become increasingly important [28,29,46,48,50]. These days, it is not enough to apply basic statistics (e.g., best, worst, average, and standard deviation) applied to some newly proposed algorithm, but a paper must make a more advanced statistical analysis, which shows that the algorithm is significantly better than the others. Unfortunately, such statistics require knowledge by the user to properly apply them, which is not as exact as calculating some simple statistics. This includes checking for those conditions, which must be fulfilled, so that the correct statistics can be applied.

One way is to use statistical tests that compare algorithms based on their performance [9–11,16,17]. Such comparative studies, independently of the research area (machine learning, stochastic optimization or some other research areas), are based on the idea of hypothesis testing [32].

Hypothesis testing, also called significance testing, is one method that can be used to test a hypothesis about a parameter in a population, using data measured in a data sample, or about the relationship between two or more populations, using

[*] Corresponding author at: Computer Systems Department, Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia.
E-mail addresses: tome.eftimov@ijs.si (T. Eftimov), peter.korosec@ijs.si (P. Korošec), barbara.korousic@ijs.si (B. Koroušić Seljak).

http://dx.doi.org/10.1016/j.ins.2017.07.015

data measured in data samples. The method starts by defining two hypotheses, the *null hypothesis* H_0 and the *alternative hypothesis* H_A. The null hypothesis is a statement that there is no difference or no effect and the alternative hypothesis is a statement that directly contradicts the null hypothesis by indicating the presence of a difference or an effect. This step is crucial, because mis-stating the hypotheses will disrupt the rest of the process. The next step is to select an appropriate *test statistic T*, which is a measurable function of a random sample that allows researchers to determine the likelihood of obtaining the outcomes if the null hypothesis is true. The level of significance α, also called the *significance level*, which is the probability threshold below which the null hypothesis will be rejected, also needs to be selected. The last step is to make a decision either to reject the null hypothesis in favor of the alternative or not to reject it. This decision can be made using two different approaches. In the first, all the possible values of the test statistic for which the null hypothesis is rejected, also called the *critical region*, are calculated using the distribution of the test statistic and the probability of the critical region, which is the level of significance α. Then the observed value of the test statistic T_{obs} is calculated according to the observations from the data sample. If the observed value of the test statistic lies within this critical region, the null hypothesis is rejected, and if not, it is not rejected. In the second approach, instead of defining the critical region, a *p*-value i.e., the probability of obtaining the sample outcome, given that the null hypothesis is true, is calculated. The null hypothesis is rejected, if the *p*-value is less than the significance level (typically 0.05 and 0.1), and if not, it is not rejected.

The result of hypothesis testing can either be correct or incorrect. Errors in the conclusion are called either *Type I* or *Type II*. A *Type I* error occurs if we reject the null hypothesis when it is true ($\alpha = P(Type\ I\ Error)$). A *Type II* error occurs if we fail to reject the null hypothesis when the alternative hypothesis is true ($\beta = P(Type\ II\ Error)$). The *power* of a hypothesis test is the probability of making the correct decision if the alternative hypothesis is true. That is the probability of rejecting the null hypothesis when the alternative hypothesis is true $(1 - \beta)$. Related to the *Type I* and *Type II* errors is the sample size estimation. It ensures enough data to keep the probabilities of *Type I* and *Type II* errors ($\alpha\, and\, \beta$) at suitable levels [32].

In order to select an appropriate statistical test, and to choose between a parametric and a nonparametric test [16], the first step is to check the assumptions of the parametric tests, also called the required conditions for safe use of parametric tests. If the data does not satisfy the required conditions for safe use of parametric tests, then the tests can result in incorrect conclusions, and it is better to use an analogous nonparametric test. In general, a nonparametric test is less restrictive than a parametric one, but it is also less powerful than a parametric one, when the required conditions for safe use of parametric test are satisfied [16].

Many statisticians [19] have also shown that researchers have difficulties performing empirical studies and this could lead to misinterpretation of the results. Levine et al. [33] also found that hypothesis testing is frequently misunderstood and abused. For example, due to the nature of stochastic optimization algorithms a set of independent runs must be executed on a single instance of a problem to get a relevant data set for which the average or the median are typically calculated. These values are then used in further statistical comparisons. The decision of using average or median, can have a great influence on the final result of the statistical test. To make things worse, even selecting either of them can have a negative outcome to the relevant results of the statistical test. For example, averaging is sensitive to outliers, which needs to be considered especially because stochastic optimization algorithms could have poor runs [3,41], while both of them are sensitive to errors inside some ϵ-neighborhood, which show that there is some statistically significant difference, even if there is none. An ϵ-neighborhood is the set of all numbers whose distance from a given number is less than some specified number ϵ. For example, the Friedman ranking scheme ranks the algorithms for each problem separately, with the best performing algorithm ranked number 1, the second best rank 2, and so on. In case of ties, average ranks are assigned. Let us suppose that we have three algorithms and their corresponding average values or medians from the multiple runs are in some ϵ-neighborhood (e.g., $2e-09,\ 3e-09,$ and $4e-09$). According to these values, the algorithms will obtain different ranks because there are no ties between these values. But the question is how to define the ϵ-neighborhood for different test functions that contain data in different ranges (e.g., $e-09,\ e-02,\ e+02,\ e+03,$ etc.). So it can happen that the averages or medians are in some ϵ-neighborhood and the distributions of multiple runs are the same so that there is no difference between the performance of the algorithms and they should obtain the same rank. Also, it can happen that that the averages or medians are in some ϵ-neighborhood, but the distributions of multiple runs are not the same suggesting a difference between the performance of the algorithms and they should obtain different ranks.

For these reasons we propose a novel approach, which removes the sensitivity of simple statistics to the data and enables the calculation of more robust statistics unaffected by outliers or by errors inside the ϵ-neighborhood. To make things even simpler, we have also created a statistical tool, which automatizes this process, so that even a user with limited statistical knowledge is able to properly apply it to the problem. Here we would like to note that we are aware of multi-modality problem. In this paper, we only focus on improving the robustness of current approaches that are used for the statistical comparison of stochastic optimization algorithms.

The remainder of the paper is organized as follows: Section 2, gives an overview of the related work. Section 3 describes the problem in depth. A new ranking scheme that will be used to compose a sample of results for each algorithm for multiple-problem analysis is presented in Section 4. Section 5 presents the experimental study of statistical comparison of stochastic optimization algorithms over multiple problems by using the new proposed ranking scheme together with a discussion of the results obtained. Section 6 presented the power analysis of different statistical tests with the proposed ranking scheme in order to see which results are more reliable. In Section 7 we present a discussion about our approach and how it differs from the common approach. The conclusions are presented in Section 8.