Neurocomputing 147 (2015) 235–238

Contents lists available at ScienceDirect

Neurocomputing

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

Letters

An alternative way of presenting statistical test results when evaluating the performance of stochastic approaches

Thomas Weise [a,*], Raymond Chiong [b]

[a] USTC-Birmingham Joint Research Institute in Intelligent Computation and Its Applications (UBRI), University of Science and Technology of China, Hefei 230027, Anhui, China

[b] School of Design, Communication and Information Technology, The University of Newcastle, Callaghan, NSW 2308, Australia

ARTICLE INFO

Article history:
Received 11 November 2013
Received in revised form
28 May 2014
Accepted 24 June 2014
Communicated by A. Abraham
Available online 9 July 2014

Keywords:
Statistical tests
Directed acyclic graph
Hasse diagram
Stochastic algorithms
Optimization

ABSTRACT

Stochastic approaches such as evolutionary algorithms have been widely used in various science and engineering problems. When comparing the performance of a set of stochastic algorithms, it is necessary to statistically evaluate which algorithms are the most suitable for solving a given problem. The outcome of statistical tests comparing $N \geq 2$ processes, where N is the number of algorithms, is often presented in tables. This can become confusing for larger numbers of N. Such a scenario is, however, very common in both numerical and combinatorial optimization as well as in the domain of stochastic algorithms in general. In this letter, we introduce an alternative way of visually presenting the results of statistical tests for multiple processes in a compact and easy-to-read manner using a directed acyclic graph (DAG), in the form of a simplified Hasse diagram. The rationale of doing so is based on the fact that the outcome of the tests is always at least a strict partial order, which can be appropriately presented via a DAG. The goal of this brief communication is to promote the use of this approach as a means for presenting the results of comparisons between different optimization methods.

1. Introduction

In the field of numerical and/or combinatorial optimization, simulation experiments are often used to determine which method is the best for solving a given problem. Broadly speaking, techniques for addressing different kinds of optimization problems can be classified into two major classes: *exact* and *stochastic* algorithms. The latter is typically called into play when the problems to be tackled are large, complex, dynamic, or involve the optimization of more than one objective function (see [5,6,9,20]).

Due to the stochastic nature of the algorithms, however, the optimization results could vary every time a particular algorithm of this class is executed. As such, it becomes mandatory to run the algorithm several times on the same problem instance and collect statistics of the results (median, interquartile range, mean, standard deviation, etc.). These statistics can only give a very rough impression of the algorithm's behavior, as pointed out by Weise et al. [23]. When comparing the performance of two or more stochastic algorithms on a problem instance, statistical tests (e.g., the Mann–Whitney U test or Wilcoxon rank-sum test, t-test, Kruskal–Wallis test, etc.) are required to claim with a certain level of confidence as to which algorithm is the best. The conclusion that can be drawn from such tests is usually something like

> "With a probability to err of no more than 0.01 (i.e., at a significance level of 1%), we can state that 'Method A' outperforms 'Method B'."

or

> "At a significance level of 5% (or with a maximally allowed type I error probability of 0.05), no statistically significant difference can be detected between the performance of 'Method A' and 'Method B'."

Instead of following the standard way of presenting statistical test results using tables, in this letter we discuss a very simple graphical representation to visualize the outcome of statistical tests used for comparing N processes (or stochastic distributions) based on datasets sampled from them. This simple approach was, to the best of our knowledge, first conceived by Burda [3], and has thereafter been adopted or independently used by several researchers in their work (e.g., [18,20–22,25]). Recently, software implementations of the approach have been made available by Burda [4] and Voigt et al. [19]. The positive aspect of the approach is the simplicity and clarity of its presentation, although there has also been reservation from some readers and reviewers about its non-standard way of representing the data. The goal of this letter is therefore to promote the use of this approach to a wider audience.

*: Corresponding author.

http://dx.doi.org/10.1016/j.neucom.2014.06.071
0925-2312/

2. An illustration of $N(N-1)/2$ comparisons

Generally, statistical tests [7,12,13,17] are tools to compare processes that produce measurable outputs, which can be represented as real numbers. Often, two such processes P_1 and P_2 are compared with the goal to find which of the two tends to produce smaller (or larger) outputs. Given finite samples (observations) of these processes, this question can be answered with a certain level of confidence by applying statistical tests such as the Mann–Whitney U test [14]. Based on a significance level α, i.e., a threshold for the highest acceptable probability to make a false statement, a significant difference between P_1 and P_2 is either confirmed or rejected.

If $N \geq 2$ processes P_1, P_2,..., P_N are observed, then the previous question can be extended to finding which of them tends to produce the smallest elements and to detect interrelations. One way to do this is to compare each process with every other process, again using the statistical test of choice. There are two issues with this procedure: (1) it requires provisions such as the conservative Bonferroni correction [8] or post hoc methods like a Nemenyi test [16] after a Friedman test [10] to avoid statistical errors[1] (see Demšar [7] or García and Herrera [11] for detailed discussions of more sophisticated statistical approaches and better recommendations); (2) it will result in (at most) $N(N-1)/2$ outcomes, which are hard to visualize. Here, we focus on the latter issue. A common way to represent the outcomes is to use a table (matrix) $T_{i,j} \in \{+, -, 0\}$. A value of $T_{i,j}=+$ in the ith row and jth column means that process P_i has significantly larger outputs than process P_j, a "$-$" stands for smaller outputs, and 0 symbolizes that no significant difference could be detected (at the given significance level α).

Table 1 shows an example of how a common tabular illustration of the comparison results for eleven processes P_1–P_{11} could look like. Only the upper triangle of the table needs to be populated since $T_{i,j}=+ \Rightarrow T_{i,j}=-$, $T_{i,j}=- \Rightarrow T_{i,j}=+$, $T_{i,j}=0 \Rightarrow T_{i,j}=0$, and $T_{i,j}=0$ for all $i, j \in 1...N$. From the example, it is clear that with the increasing number of processes, it becomes more difficult to recognize the order of the processes according to the tests from such a table.

3. Graph-based notation

3.1. An example

Clearly, a full set of $N(N-1)/2$ test results defines a partial order on the compared processes. Besides using a table or matrix, such a partial order can be illustrated in the form of a directed acyclic graph (DAG), as sketched in Fig. 1(a). Such graphical representations of partial orders are known as the Hasse diagrams [1,2] and have been used in the area of education [25]. In our case, each process can be represented as a node in a graph. Here, $T_{i,j}=+$ will result in a directed edge from the node labeled with P_j to the node labeled with P_i. A "$-$" results in a directed edge into the opposite direction and a "0" is represented by having no edge between the corresponding nodes.

Since the test results form a transitive order, edges that are sufficiently explained by transitivity can be omitted in the graph (and actually, the corresponding tests do not need to be performed in the first place). Hence Fig. 1(a) does not contain an arrow from node P_2 to P_1, since that one is already subsumed by the arrow from P_2 to P_{11} and from P_{11} to P_1. The graph sketched in Fig. 1(a) is easier to read than Table 1. The Hasse diagram-based notion can be further simplified by combining those nodes for which all incoming arrows come from the same origins and all outgoing arrows target the same nodes.

[1] The first author noted that he did not take such measures in his previous work due to ignorance of the issue.

Fig. 1(b) represents such a simplification. It is our strong belief that this representation could be a good alternative to the tabular representation, because of its compactness, clarity, and ease of use. From Fig. 1(b), it can immediately be seen that processes P_1 and P_7 tend to have the largest outputs while P_4 has the smallest. There is no significant difference between P_9 and P_6 or P_3, but P_9 tends to produce smaller outputs than P_{10}. The outputs of P_5 tend to be larger than those of P_3, but there are no significantly differences from those of P_2.

3.2. Formal definition

Given a set P of N processes P_i: $i \in 1...N$ and a statistical test result matrix $T_{i,j} \in \{+, -, 0\} \forall i, j \in 1...N$, the graph-based representation G is defined as follows:

1. For each $P_i \in P$, there exists exactly one node labeled with P_i in G.
2. A node may be labeled with a set S of multiple process names if and only if $\forall P_i, P_j \in S \Rightarrow (\forall P_k \in P \Rightarrow T_{i,k}=T_{j,k}$ and $T_{k,i}=T_{k,j})$ holds.
3. There exists a directed edge from the node labeled with P_j to the node labeled P_i if and only if
   (a) $T_{i,j}=+$ (and, hence, $T_{j,i}=-$) and
   (b) $\neg \exists P_k \in P$: $(T_{i,k}=+) \wedge (T_{k,j}=+)$.

The graph can be created by using existing tools such as those of Burda [4] and Voigt et al. [19]. Alternatively, one can first create a graph that contains a directed edge for each $T_{i,j}=+$. This graph can then be iteratively simplified by deleting edges for which rule 3 above holds and merging nodes according to rule 2 until further reduction is possible. Since the manual layout of larger graphs is tedious, the resulting graph could be represented in a text-based format like the DOT language, which then can be rendered by tools such as Graphviz (see http://www.graphviz.org/).

3.3. How to use

We want to emphasize that a diagram such as Fig. 1 should always be accompanied by a descriptive note stating the applied test and the test's configuration, the significance level, and the meaning of the presence of a directed edge in the graph. An example for this notion could be

> "Fig. 1(b) shows the outcome of the application of a two-tailed Mann–Whitney U test with the Bonferroni correction and a significance level of 1% (type I error probability ≤ 0.01) to the data sampled from processes P_1 to P_{11}. A directed edge from a node P_i to a node P_j means that, according to the applied test, P_i produces {larger/smaller/better} outcomes than P_j."

Such a description text is not longer than what would be needed to properly define the meaning of the tabular result expression (see the example in Table 1).

4. Other visualization techniques

Before we end, it is worth pointing out that there exist several other visualization techniques for illustrating statistical test results. However, these techniques may quickly get harder to read once the number of compared datasets increases.

One of these visualization techniques is notched boxplots as described by McGill et al. [15]. Boxplots represent data. They do not represent statistical test results. However, if the notches of two boxes representing different datasets do not overlap, this is an indicator that their medians may be significantly different at a 5% error level. See Wickham and Stryjewski [24] for more discussion on variants of boxplots.