An Alternative Way of Presenting Statistical Test Results when Evaluating the Performance of Stochastic Approaches

Thomas Weise*a, Raymond Chiongb

*aUSTC-Birmingham Joint Research Institute in Intelligent Computation and Its Applications (UBRI)
University of Science and Technology of China, Hefei 230027, Anhui, China
bSchool of Design, Communication and Information Technology
The University of Newcastle, Callaghan, NSW 2308, Australia

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.

KEY WORDS: statistical tests; directed acyclic graph; Hasse diagram; stochastic algorithms; optimization

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 Engelbrecht A.P. (2007), Chiong R. (2009), Weise T. (2009a), Chiong R. et al. (2012)).

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 T. et al. (2014). 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 M. (2006), and has thereafter been adopted or independently used by several researchers in their work (e.g., Weise T. (2009a), Weise T. (2009b), Zoubek L. and Burda M. (2009), Toledo C.F.M. et al. (2011), Weise T. and Tang K. (2012)). Recently, software implementations of the approach have been made available by Burda M. (2013) and Voigt K. et al. (2013). 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.

2. **AN ILLUSTRATION OF \( N(N-1)/2 \) COMPARISONS**

Generally, statistical tests (Sheskin D.J. (2004), Harlow L.L. et al. (1997), Levin J.R. (1998), Demšar J. (2006)) 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 (Mann H.B. and Whitney D.R. (1947)). Based on a significance level \( \alpha \), 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 \( P_1, P_2, \ldots, P_N \) 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 (Dunn, O.J., 1961) or post hoc methods like a Nemenyi (1961) test after a Friedman (1937) test to avoid statistical errors¹ (see Demšar J. (2006) or García S. and Herrera F. (2008) 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) \( T_{ij} \in \{+, -, 0\} \). A value of \( T_{ij} = + \) in the \( i \)th row and \( j \)th 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 \( \alpha \)).

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

¹ The first author noted that he did not take such measures in his previous work due to ignorance of the issue.
3. GRAPH-BASED NOTATION

An Example

Clearly, a full set of $\mathcal{O}(N^2)$ 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 Figure 1(a). Such graphical representations of partial orders are known as Hasse diagrams (Birkhoff G. (1948), Baker K.A. et al. (1972)) and have been used in the area of education (Zoubek L. and Burda M. (2009)). In our case, each process can be represented as a node in a graph. Here, $T_{ij} = +$ 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 Figure 1 does not contain an arrow from node $P_3$ to $P_1$, since that one is already subsumed by the arrow from $P_2$ to $P_1$ and from $P_3$ to $P_1$. The graph sketched in Figure 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.

Figure 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 Figure 1(b), it can immediately be seen that processes $P_7$ and $P_9$ tend to have the largest outputs while $P_4$ has the smallest. There is no
significant difference between $P_9$ and $P_3$ or $P_6$, but $P_9$ tends to produce smaller outputs than $P_{10}$. The outputs of $P_3$ tend to be larger than those of $P_2$, but there are no significantly differences from those of $P_2$.

Formal Definition

Given a set $P$ of $N$ processes $P_i : i \in 1...N$ and a statistical test result matrix $T_{ij} \in \{+,-,0\} \forall i,j \in 1...N$, 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 $\forall P_i, P_j \in S \Rightarrow (\forall P_k \in P \Rightarrow T_{ik} = T_{jk} \text{ and } T_{ki} = T_{kj})$ 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. $T_{ij} = +$ (and, hence, $T_{ji} = -$) and
   b. $\neg \exists P_k \in P : (T_{ik} = +) \land (T_{kj} = +)$.

The graph can be created by using existing tools such as those of Burda, M. (2013) and Voigt K. et al. (2013). Alternatively, one can first create a graph that contains a directed edge for each $T_{ij} = +$. 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/).

How to Use

We want to emphasize that a diagram such as Figure 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:

“Figure 1(b) shows the outcome of the application of a two-tailed Mann-Whitney $U$ test with Bonferroni correction and a significance level of 1% (type I error probability $\leq 0.01$) 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 of 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 R. Tukey J.W., and Larsen W.A. (1978). 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 H. and Stryjewski L. (2011) for more discussion on variants of boxplots.
Another possible way of visualization is, instead of printing a table with the win/loss/undecided test results in the cells one can print a table where the cells are colored according to the \( p \)-value returned by the tests, e.g., light gray = low \( p \)-value, black = high \( p \)-value. The advantage of doing this may be that it would allow us to create really large tables, since it uses pixel colors instead of text.

A nice alternative visualization technique can be found in the work of Demšar J. (2006), in which all algorithms in comparison are listed on an axis denoting their average ranks. Algorithms with results that are not statistically different are connected with lines.

Line and letter diagrams are similar to Demšar’s approach, in that they connect groups of datasets that are not significantly different. However, some examples as shown in Burda M. (2013) indicate that they may become complicated even for small \( N \), and that it may not always be possible to see the statistically significant differences of the results at first glance. Burda further hints that the Hasse diagram based approach is better in this respect.

5. **CONCLUDING REMARKS**

We ourselves are not statisticians but mere users of simple statistical tools for analyzing the experimental outcomes. The presented approach to visualizing the outcome of statistical tests is the result of hands-on attempts to present data for scientific papers with limited space. Our experience with this notation is positive, so we wish to promote its use in the analysis of stochastic algorithms. Even though the presented approach appears to be suitable and easy-to-use for us, others may think otherwise. We would thus be very thankful for any comments, corrections, and suggestions regarding this subject.

**REFERENCES**


