Information Sciences 178 (2008) 2870–2879

Contents lists available at ScienceDirect

Information Sciences

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

A general framework for statistical performance comparison of evolutionary computation algorithms

David Shilane [a], Jarno Martikainen [b,*], Sandrine Dudoit [a], Seppo J. Ovaska [b]

[a] Division of Biostatistics, School of Public Health, University of California, Berkeley, CA, USA
[b] Faculty of Electronics, Communications, and Automation, Helsinki University of Technology, Otakaari 5 A, P.O. Box 3000, 02015 Espoo, Finland

ARTICLE INFO

Article history:
Received 18 January 2006
Received in revised form 18 February 2008
Accepted 5 March 2008

Keywords:
Evolutionary computation
Genetic algorithms
Performance comparison
Statistics
Twofold sampling
Bootstrap
Multiple hypothesis testing

ABSTRACT

This paper proposes a statistical methodology for comparing the performance of evolutionary computation algorithms. A twofold sampling scheme for collecting performance data is introduced, and these data are analyzed using bootstrap-based multiple hypothesis testing procedures. The proposed method is sufficiently flexible to allow the researcher to choose how performance is measured, does not rely upon distributional assumptions, and can be extended to analyze many other randomized numeric optimization routines. As a result, this approach offers a convenient, flexible, and reliable technique for comparing algorithms in a wide variety of applications.

1. Introduction

Evolutionary algorithms (EAs) [1,9] are used to estimate the solution to difficult optimization problems. EAs are often hand-crafted to meet the requirements of a particular problem because no single optimization algorithm can solve all problems competitively [19]. When alternative algorithms are proposed, their relative efficacies should be assessed. Because EAs follow a stochastic process, statistical analysis is appropriate for algorithm comparison. This paper seeks to provide a general methodology for comparing the performance of EAs based on statistical sampling and hypothesis testing.

Prior research in the statistical design and analysis of EAs has considered a variety of approaches. Based upon a large number of experimental trials, Penev and Littlefair [12] demonstrate that the Free Search algorithm improves upon previous results from a variety of stochastic competitors on several optimization problems. This comparison consists of defining a number of performance metrics and computing average values for each algorithm. However, like many other evolutionary computation studies, these results are not statistically analyzed and substantiated. Because of the large sample size and clear observed differences in their results, we have no reason to doubt the specific findings of the Free Search study. Indeed, a statistical analysis of these data would likely add weight to the conclusions. In proposing a general framework for statistical performance comparison of EAs and similar randomized optimization algorithms, we seek to provide an experimental framework in which the results of similar studies may be assessed according to appropriate statistical tests.

Christensen and Wineberg [3] explain the use of appropriate statistics in artificial intelligence and propose non-parametric tests to verify the distribution of an EA's estimate of a function's optimal value. Flexer [8] proposes general guidelines for

* Corresponding author.
E-mail address: martikainen@iki.fi (J. Martikainen).

0020-0255/$ - see front matter © 2008 Published by Elsevier Inc.
doi:10.1016/j.ins.2008.03.007

statistical evaluation of neural networks that can also be applied to EAs. Although a variety of non-parametric tests are available, these procedures are often limited to specific parameters of interest. For instance, the Mann–Whitney test (also called Wilcoxon's *rank sum test* [16]) may be used to assess the equality of two populations' medians without requiring any information about the data's distribution. However, such a test is not easily adapted to other parameters, such as the mean difference between the two populations, the simultaneous comparison of more than two populations at once, or a simultaneous test of both the median and another parameter of interest. Czarn [4] discuss the use of the analysis of variance (ANOVA) in comparing the performance of EAs. Similarly, Castillo-Valdivieso et al. [2] and Rojas et al. [17] employ ANOVA methods to optimize the parameter values in the design of improved EAs for specific optimizations, whereas François and Lavergne [10] rely upon a generalized linear model. However, these procedures all require distributional assumptions that are not necessarily valid and also limit the class of performance metrics that can be used. Because EAs produce results according to complex stochastic processes, often very little is known about the distribution of results across algorithmic trials. We seek to address this problem by relying solely on empirical data generated from repeated trials of competing EAs. The proposed methodology employs a bootstrap-based multiple hypothesis testing framework [6,5,15,13] that may be applied to any parameter of interest, number of simultaneous hypotheses, and data distribution. The resulting procedure establishes an experimental framework in which EAs may be compared based upon empirical data.

An EA's initial population (Section 2) consists of a set of starting values for the evolution process. Most previous EA performance comparisons have only considered results for a single initial population or even provided different inputs for each algorithm studied. Supplying different single inputs to each EA may result in a *founder effect*, in which a population's initial advantage is continually propagated to successive generations. Furthermore, relying upon a single choice of initial population can at best determine the plausibility of preferring one candidate EA to another given suitable initial conditions. We can alleviate these issues by assessing relative performance over each of a representative sample of initial populations.

For each particular initial population sampled, two EAs may be compared by testing the null hypothesis of equal performance according to a specified performance metric. Student's t-statistics [11] are commonly used to test the equality of two population means. However, the parametric t-test assumes that the data are normally distributed. If this assumption is not valid, the resulting inference may not be meaningful. Therefore, we require a more general and objective framework for statistical performance comparison of EAs.

Because we are proposing a scientific method for performance comparison, it is important to design an effective experiment that specifies how data are collected and analyzed. To collect data, we propose a twofold sampling scheme to perform repeated EA trials at each of a representative sample of possible inputs. The candidate EAs' efficacies are then assessed in a multiple hypothesis testing framework that relies upon bootstrap resampling [6,5,15,13] to estimate the joint distribution of the test statistics. This methodology establishes a procedure for EA comparison that can be considered general in the following aspects: First, the results do not rely heavily on a single advantageous input. Second, the bootstrap-based testing procedure is applicable to any distribution and requires no *a priori* model assumptions. Finally, this methodology can be applied to essentially any function of the data collected, so the researcher is free to choose how performance should be evaluated. The result is a general framework for performance comparison that may be used to compare EAs or other stochastic optimization algorithms based upon empirical data.

The paper is organized as follows: Section 2 provides a brief introduction to EAs and presents a twofold sampling scheme for data collection. Section 3 places performance comparison in a multiple hypothesis testing framework. Section 4 shows how to use the bootstrap to estimate the test statistics' underlying distribution. Section 5 introduces a variety of multiple testing procedures. Section 6 provides an example comparing the performance of two EAs seeking to minimize Ackley's function. Section 7 discusses further applications of statistics in EA performance comparison and concludes the paper.

2. Evolutionary algorithms and data collection

An EA's *cost* (or *objective*) *function* is a map $f : \mathbb{R}^D \to \mathbb{R}$ to be optimized. Any candidate solution is specified by an *individual* with a vector of *genes* (or *traits*, used interchangeably) $\mathbf{y} = (y_1, \ldots, y_D)$. Each individual has a corresponding cost given by $f(\mathbf{y})$. Given a *population* of individuals, an EA uses *evolutionary mechanisms* to successively create *offspring*, or new individuals. The evolutionary mechanisms often consist of some combination of selection, reproduction, and mutation operators. The *selection* mechanism ranks individuals by cost, determines which individuals shall produce offspring, and assigns individuals to *mating groups*. Given a mating group, *reproduction* combines the genes of individuals within the mating group one or more times to produce offspring. Finally, the *mutation* mechanism randomly alters the genetic profile of offspring immediately following conception.

An EA's *initial population* (or *input*, used interchangeably) is a set of individuals that serve as starting values for the algorithm, and its *result* is given by the minimum observed cost among all individuals produced in G generations. Once the evolutionary mechanisms are specified, one ordered iteration of these processes in sequence is considered one *generation*, and the *evolution* process proceeds for a user-specified number of generations $G \in \mathbb{Z}^+$.

An EA's result is determined by a stochastic process with two sources of variation: the initial cost and the algorithm's improvements to this cost produced by G generations of the random evolution process. Because an EA's result depends both on its initial cost and its efficacy given this initial population, a sample of result data should be collected in a *twofold sampling* scheme: we first generate a representative sample of initial populations, and then, for each of these inputs, we perform a