Applied Soft Computing 55 (2017) 265–275

Contents lists available at ScienceDirect

Applied Soft Computing

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

The impact of Quality Indicators on the rating of Multi-objective Evolutionary Algorithms

Miha Ravber, Marjan Mernik, Matej Črepinšek *

Faculty of Electrical Engineering and Computer Science, University of Maribor, Slovenia

ARTICLE INFO

Article history:
Received 17 October 2016
Received in revised form 12 January 2017
Accepted 25 January 2017
Available online 4 February 2017

Keywords:
Multi-objective optimization
Evolutionary Algorithms
Quality Indicator
Performance assessment
Chess rating

ABSTRACT

Evaluating and comparing multi-objective optimizers is an important issue. But, when doing a comparison, it has to be noted that the results can be influenced highly by the selected Quality Indicator. Therefore, the impact of individual Quality Indicators on the ranking of Multi-objective Optimizers in the proposed method must be analyzed beforehand. In this paper the comparison of several different Quality Indicators with a method called Chess Rating System for Evolutionary Algorithms (CRS4EAs) was conducted in order to get a better insight on their characteristics and how they affect the ranking of Multi-objective Evolutionary Algorithms (MOEAs). Although it is expected that Quality Indicators with the same optimization goals would yield a similar ranking of MOEAs, it has been shown that results can be contradictory and significantly different. Consequently, revealing that claims about the superiority of one MOEA over another can be misleading.

1. Introduction

Multi-objective optimization is an area that deals with optimization of multi-objective optimization problems (MOPs). A multi-objective optimization problem can be defined as:

$$\begin{aligned} &\text{Minimize } F(x) = (f_1(x), \ldots, f_m(x)) \\ &\text{Subject to} \\ &\quad g_i(x) \geq 0 \qquad i = 1, 2, \ldots, p \\ &\quad h_i(x) = 0 \qquad i = 1, 2, \ldots, q, \end{aligned} \tag{1}$$

where $x = (x_1, x_2, \ldots, x_n) \in X$ is an n-dimensional decision vector bounded in the decision (variable) space X, and the objective function vector $F : X \to \mathbb{R}^m$ consists of m ($m \geq 2$) real-valued objective functions f_i and $\mathbb{R}^m$ is the objective space. p is the number of inequality constraints, while q is the number of equality constraints.

Let $x, y \subset X$ denote two solutions. x is said to dominate y ($f(x) \prec f(y)$) if for each objective i is $f_i(x) \leq f_i(y)$ and at least one objective j is $f_j(x) < f_j(y)$. A solution x^* is called a Pareto optimal solution if there exists no solution $x \subset X$ such that $f(x) \prec f(x^*)$. The set of these optimal solutions is termed as the Pareto front (PF) in the objective space $\mathbb{R}^m$ and the Pareto set (PS) in the decision space X. The set of optimal solutions in the PF represent the best trade-offs between different objectives, since a single solution cannot optimize all objectives simultaneously. Thus, the goal of Multi-Objective Optimization (MOO) is to obtain the Pareto optimal front. Since many Multi-Objective Optimization problems are difficult to solve, the outcome of the optimization is usually an approximation of the Pareto front. These approximations need to be evaluated in order to compare them. Evaluating the quality of these approximations is itself an MOP. Zitzler et al. [1] suggested three optimization goals (aspects) that need to be measured:

- The distance of the resulting approximation set to the Pareto optimal front should be minimized (convergence).
- A good (in most cases uniform) distribution of the solutions found is desirable (uniformity).
- The extent of the obtained approximation front should be maximized (spread).

Comparing the performance of MOEAs remains an open problem. The most popular measures are Quality Indicators (QI); the term "performance metric" is also used to quantify the differences between approximation sets.

Many different QIs for measuring the quality of approximation sets have been proposed in the literature [1–10]. Each QI has been designed with a standpoint that takes one or more previously mentioned optimization goals into consideration. This means that no single indicator alone can reliably measure all optimization goals.

* Corresponding author.
E-mail addresses: miha.ravber@um.si (M. Ravber), marjan.mernik@um.si (M. Mernik), matej.crepinsek@um.si (M. Črepinšek).

http://dx.doi.org/10.1016/j.asoc.2017.01.038
1568-4946/

It should be noted that several surveys and experiments have been conducted to analyze individual QIs [4,8,11,12,13]. The results have shown inconsistencies in the assessment of various approximation sets. This was also confirmed by our results, which revealed further surprising dimensions of the impact of QIs on the rankings of MOEAs and that the rankings can be contradictory even between QIs with the same optimization goals. It was argued in [11] and [12], that, without established comparison criteria, claims based on heuristically chosen QIs do little to determine a given MOEA's actual efficiency and effectiveness. In addition, the conclusions are useless for answering the question of which algorithms are superior to others, so can it be argued that one algorithm is better than another even though the outcome depends on the selected QI?

The aim of this paper is to acquire new knowledge about QIs and to obtain a better insight into the impact of the selected QI in the comparison of MOEAs using a Chess Rating System for Evolutionary Algorithms (CRS4EAs) [14]. In CRS4EAs each player's (MOEA) performance is represented by rating obtained with the Glicko-2 rating system. Based on the differences in ratings of MOEAs, significant differences can be detected. Using different Quality Indicators in the comparison we can observe their impact on MOEA's rating and consequently on significant differences amongst MOEAs. The main contributions of this paper are: A detailed analysis of QIs using a novel method called CRS4EAs, and finding coherent and robust QIs to increase the reliability of the assessment of Pareto approximation sets and, thereby, also increase the legitimacy of claims about MOEA performance.

This paper is an extended version of the conference paper published in [15], where a CRS4EAs was first introduced for comparing MOEAs. We extended our previous work with a more detailed analysis of QIs and added a second scenario where we repeat the experiment on a real world problem to increase the practical value of the research. This will give us an even better insight on the impact of QIs on the rating of MOEAs.

The remainder of the paper is organized as follows. Section 2 introduces some basic concepts of Quality Indicators. The CRS4EAs is presented in Section 3. Section 4 presents, the execution of the experiment and results. Finally, the paper concludes in Section 5.

2. Quality Indicators

Solutions of MOPs can be compared using dominance relations [4]. However, there are numerous limitations in using this approach. For example, the extent to which one approximation is better than another cannot be expressed nor can it be expressed in which aspects this is so. Furthermore, when using dominance relations, there are cases in which approximation sets are incomparable. This applies particularly for problems with larger number of objectives caused by the phenomena known as the curse of dimensionality [16]. QIs have been designed in order to overcome these limitations. These QIs measure approximations of Pareto optimal fronts quantitatively. Therefore, QIs, are in essence, functions that assign each approximation set a real number that reflects different aspects of quality or quality differences. Zitzler et al. [4] defined a Quality Indicator I as an m-ary function $I : \Omega^m \to \mathbb{R}$ that assigns each vector $(A_1, A_2, \ldots, A_m)$ of m approximation sets a real value $I(A_1, \ldots, A_m)$ where Ω is the set of all approximation sets. Once the approximation sets are evaluated by QIs, different conclusions can be drawn about their relations. For different aspects of quality, different QIs need to be used.

Quality Indicators have been categorized into different groups from different points of view to understand their nature better [4,11,2]. They are categorized mainly by the aspects of quality that they assess. These aspects include the closeness to the Pareto-optimal front, the number of elements of the Pareto-optimal front found, and the maximum spread of solutions. Quality Indicators are also classified based on the number of approximation sets they take as an argument. Unary QIs accept one approximation and binary accept two. However, in principle, QIs that accept an arbitrary number of arguments are also possible. When evaluating with unary QIs the resulting real values need to be compared in order to see which result set is better. Binary QIs, in contrast, compare two result sets to determine which one is better. Therefore, when comparing t algorithms using binary QIs, $t(t\text{ -}1)$ distinct indicator values are obtained. Some unary QIs require a reference set to perform the evaluation, which must be taken into consideration since real-world problems have unknown Pareto-optimal fronts. When the reference set is available, any QI can be converted from binary to unary. There are also other categories that are not used as often, such as computational complexity, the sensitivity to scaling, the number of objectives, etc. It is also desirable that a QI be compatible and complete with respect to dominance relations. A Quality Indicator I is compatible with the dominance relation if and only if $\forall A, B \in \Omega$, $I(A)$ is better than $I(B) \to A$ dominates B. A Quality Indicator I is complete with the dominance relation if and only if $\forall A$, $B \in \Omega$, A dominates $B \to I(A)$ is better than $I(B)$ [4].

Quality Indicators need interpretation, and different comparison methods can be used. This is best illustrated by Zitzler (Fig. 1) [4] where concepts of comparison methods are presented using either only unary or only binary QIs. Case (a) uses a single unary QI, (b) a single binary QI, and (c) a combination of two unary QIs. In cases (a) and (b), the indicator I evaluates the approximation sets A and B. The result is passed to the interpretation function E which returns true if the first approximation is better than the second. In case (c), two unary indicators are applied to A and B then the resulting two indicator values are combined in a vector, $I(A)$ for A and vector $I(B)$ for B. The vectors are passed to the interpretation function E that decides the outcome. The interpretation function returns true if the first approximation (A) is better than the second (B), otherwise it returns false.

In this paper, eleven QIs (Coverage of two sets (CS), additive ϵ Indicator ($I_{\epsilon+}$), Generational Distance (GD), Hypervolume (HV), Inverted Generational Distance (IGD), improved IGD (IGD+), Maximum Pareto Front Error (MPFE), MaximumSpread (MS), the R2 indicator (R2), Spacing (S) and generalized spread (Δ)) are used, based on their prevalence in literature and different properties [17–19]. Selected Indicators are listed with their characteristics in Table 1. *CS* is one of the few commonly used binary QI in literature. Its advantage is that it does not require a reference set. As we can see, this is a shortcoming for most of unary QIs. As already mentioned, they are categorized by aspects of quality that they assess. Some of them even fall in all three categories. In a fair comparison we want to evaluate all aspects of quality and we therefore ask ourselves why not use just these QIs? As we have already stated, evaluating the quality of approximation sets is an MOP. This means that a single QI can not reliably assess all aspects at once. This was also demonstrated in our experiments.

When comparing algorithms, usually a handful of QIs are selected and then the experiment is performed and evaluated with selected statistical methodologies. The CRS4EAs was used in our case. The outcome of a comparison in CRS4EAs was determined by methods a) and b) (Fig. 1), depending whether the QI was unary or binary.

3. Chess Rating System for Evolutionary Algorithms (CRS4EAs)

In order to analyze QIs we observed their impact on the rating of MOEAs using the method called CRS4EAs [14]. With CRS4EAs we can quantify the performance differences between MOEAs and