Information Sciences 396 (2017) 185–201

Contents lists available at ScienceDirect

Information Sciences

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

Analyzing evolutionary optimization and community detection algorithms using regression line dominance

Anupam Biswas*, Bhaskar Biswas

Department of Computer Science and Engineering, Indian Institute of Technology (BHU), Varanasi, India

ARTICLE INFO

Article history:
Received 15 June 2016
Revised 2 February 2017
Accepted 17 February 2017
Available online 20 February 2017

Keywords:
Evolutionary optimization algorithms
Linear regression
Particle swarm optimization
Differential evolution
Visual analysis
Community detection algorithms

ABSTRACT

In this paper, a visual analysis methodology is proposed to perform comparative analysis of guided random algorithms such as evolutionary optimization algorithms and community detection algorithms. Proposed methodology is designed based on quantile-quantile plot and regression analysis to compare performance of one algorithm over other algorithms. The methodology is extrapolated as one-to-one comparison, one-to-many comparison and many-to-many comparison of solution quality and convergence rate. Most of the existing approaches utilize both solution quality and convergence rate to perform comparative analysis. However, the many-to-many comparison i.e. ranking of algorithms is done only with solution quality. On the contrary, with proposed methodology ranking of algorithms is done in terms of both solution quality and convergence rate. Proposed methodology is studied with four evolutionary optimization algorithms on 25 benchmark functions. A non-parametric statistical analysis called Wilcoxon signed-rank test is also performed to verify the indication of proposed methodology. Moreover, methodology is also applied to analyze four state-of-the-art community detection algorithms on 10 real-world networks.

1. Introduction

Evolutionary Optimization Algorithms (EOAs) are guided random algorithms [2,38,53,54,60]. EOAs are random in nature but mostly they take inspiration from natural phenomena to guide the algorithm towards optima. Often EOAs are referred as heuristic stochastic processes. EOAs change their current state to next state with the strategy adopted, while maintaining some degree of randomness to ensure exploration of solution space. EOAs are widely used in various applications of different domains mainly wherever require optimization [8,10,17,59]. With the rapid growth of applications of EOAs in several application domains, it is necessary to develop effective evaluation methodology that helps in understanding when the algorithm performs better or best suited problems for the algorithm.

Most of the existing evaluation techniques primarily focus on solution quality of EOAs, which include non-parametric approaches [12,15,42], parametric approaches [11,46], Bootstrapping [9,45], Exploratory Landscape Analysis (ELA) [39] and drift analysis [24] and so on. Among these approaches, non-parametric and parametric analysis are comparatively easier and utilized immensely to evaluate EOAs. Non-parametric approaches estimate different parameters against the solutions obtained over multiple executions of an algorithm. On the other hand, parametric approaches infer parameters from the probability

* Corresponding author.
E-mail addresses: abanumail@gmail.com (A. Biswas), bhaskar.cse@iitbhu.ac.in (B. Biswas).

http://dx.doi.org/10.1016/j.ins.2017.02.050

distribution of solutions instead of using directly different solutions. Recently, these approaches have been improved further to choose efficient parameter settings and to analyze performance in terms of specific parameter [14,47].

From the perspective of applications, good quality solutions are essential for better performance of the application. However, the cost incurred by the EOAs to obtain solutions is unavoidable and has significant impact on application in terms of efficiency. Generally, convergence rate is analyzed to ensure minimal cost. Since EOAs are iteratively guided towards the optima, in each iteration improvement in the solution quality is tracked. The EOAs reaching optimal solution fast are considered as efficient algorithms. Convergence rate analysis is an visual inspection method to determine efficiency of algorithms. Major drawback of existing convergence analysis method is that the EOAs cannot be ranked. Nevertheless, the performance of one algorithm can be compared with other EOAs easily. In recent years, visual analysis approach [37,55] and mathematical approaches [23,31,43,56] have drawn attention to evaluate EOAs. Motivated by the famous idiom "A picture is worth a thousand words" [13,25], a simple regression line dominance mechanism has been developed in this paper to analyze EOAs visually.

The proposed analysis methodology has been explored in three directives such as one-to-one comparison, one-to-many comparison and for ranking, many-to-many comparison. The methodology is designed based on the dominance properties of data points in quantile-quantile plot. A shifting mechanism has been developed using simple regression analysis of data points in quantile-quantile plot. The preservation of dominance of data points in regression line is analyzed theoretically and derived important properties of regression line dominance. Proposed approach analyzes the performance of EOAs in terms of both solution quality and convergence rate. In earlier works we have analyzed EOAs only in terms of solution quality using regression line dominance mechanism [3,5]. This work significantly extends the earlier work as well as results with discussions more in-depth. The regression line dominance mechanism is also extended for convergence rate and analyzed with best performing EOAs reported in CEC 2013 Special Session & Competition on Real-Parameter Single Objective Optimization [34]. A comparative analysis with other non-parametric test is also included. Further more, the methodology is also used to evaluate community detection algorithms, which is an application of EOAs in social network domain.

Rest of the paper is organized as follows: Section 2 provides detail about point dominance and regression line dominance with suitable example, Section 3 elaborates regression line shifting and derives important properties, Section 4 describes comparative analysis methodology of solution quality and convergence in detail with regression line dominance mechanism, Section 5 studies the proposed analysis methodology with 25 benchmark functions and details experimental setups, Section 6 analyzes four community detection algorithm using proposed methodology and finally, Section 7 concludes with the remarks about the advantages and drawbacks of proposed methodology.

2. Dominance mechanism

2.1. Point dominance

Consider two sets of data $D_a = \{a_i | i = 1, 2, 3, \ldots, r\}$ and $D_c = \{c_i | i = 1, 2, 3, \ldots, r\}$. The set whose dominance is to be evaluated is referred as *actor*, while the set with respect to whom the dominance is evaluated is referred as *competitor*. We consider equal number of items in both actor and competitor. We use quantile-quantile plot that is simply a plot of sorted D_a versus D_c. We specify x-axis for plotting actor and y-axis for plotting competitor. Each point (a_i, c_i) in the plot is a pair of $a_i \in D_a$ and $c_i \in D_c$. Point dominance at any point (a_i, c_i) is defined in reference to a neutral line.

Definition 2.1 (Neutral Line (NL))**.** A line $Y = mX + c$ is referred as NL if it goes through the origin (i.e. intercept $c = 0$) and gradient $m = 1$. The line is neutral in the sense that, at every point (x, y) in the NL $x = y$. Thus, simply we say $X = Y$ line is a NL.

We interpret larger value implies higher the dominance.[1] In reference to the NL, we define three kinds of possible dominance at any point (a_i, c_i) in the quantile-quantile plot as follows:

- **Non-dominance:** Dominance of actor over competitor at any point (a_i, c_i) that *lies on* the NL is referred as non-dominance since $a_i = c_i$ i.e. neither a_i dominates c_i nor c_i dominates a_i.
- **Actor-dominance:** Dominance of actor over competitor at any point (a_i, c_i) that *lies above* the NL is referred as actor-dominance since $a_i > c_i$ i.e. a_i dominates c_i.
- **Competitor-dominance:** Dominance of actor over competitor at any point (a_i, c_i) that *lies below* the NL is referred as actor-dominance since $a_i < c_i$ i.e. a_i is dominated by c_i.

In Fig. 1, points P_1, P_2 and P_3 of the quantile-quantile plot are examples of non-dominance, actor-dominance and competitor-dominance respectively. *Why quantile-quantile plot?* There are several advantages of quantile-quantile plot over simple scatter plot. Quantile-quantile plot maintains non-decreasing order while plotting data so smaller values are paired with smaller one while larger values with larger one. In case of scatter plot, mostly smaller values are paired with larger one and vice-versa. Therefore, scatter plot is unsuitable for defining dominance. An example scatter plot is shown in Fig. 1 considering same data that are used in Fig. 2. Clearly, in scatter plot almost half of the points are competitor-dominant. However, for the same data, in quantile-quantile plot most points were actually actor-dominant. Besides, quantile-quantile plot

[1] We will interpret smaller value implies higher the dominance later on when deal with minimization problems.