Journal of Network and Computer Applications ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

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

Evolutionary optimization: A big data perspective

Maumita Bhattacharya [a,*], Rafiqul Islam [a], Jemal Abawajy [b]

[a] School of Computing and Mathematics, Charles Sturt University, Australia
[b] School of Information Technology, Deakin University, Australia

ARTICLE INFO

Article history:
Received 6 January 2014
Received in revised form
23 June 2014
Accepted 21 July 2014

Keywords:
Optimization
Big data
Evolutionary algorithm
High dimensionality
Non-linearity

ABSTRACT

Stochastic search techniques such as evolutionary algorithms (EA) are known to be better explorer of search space as compared to conventional techniques including deterministic methods. However, in the era of big data like most other search methods and learning algorithms, suitability of evolutionary algorithms is naturally questioned. Big data pose new computational challenges including very high dimensionality and sparseness of data. Evolutionary algorithms' superior exploration skills should make them promising candidates for handling optimization problems involving big data. High dimensional problems introduce added complexity to the search space. However, EAs need to be enhanced to ensure that majority of the potential winner solutions gets the chance to survive and mature. In this paper we present an evolutionary algorithm with enhanced ability to deal with the problems of high dimensionality and sparseness of data. In addition to an informed exploration of the solution space, this technique balances exploration and exploitation using a hierarchical multi-population approach. The proposed model uses informed genetic operators to introduce diversity by expanding the scope of search process at the expense of redundant less promising members of the population. Next phase of the algorithm attempts to deal with the problem of high dimensionality by ensuring broader and more exhaustive search and preventing premature death of potential solutions. To achieve this, in addition to the above exploration controlling mechanism, a multi-tier hierarchical architecture is employed, where, in separate layers, the less fit isolated individuals evolve in dynamic sub-populations that coexist alongside the original or main population. Evaluation of the proposed technique on well known benchmark problems ascertains its superior performance. The algorithm has also been successfully applied to a real world problem of financial portfolio management. Although the proposed method cannot be considered big data-ready, it is certainly a move in the right direction.

1. Introduction

Many real life problems often involve non-linear, high-dimensional, complex search space that may be riddled with many local minima or maxima. These issues get incredibly more complicated in problems involving big data. Stochastic methods, such as evolutionary algorithms perform a more exhaustive search of the model space as compared to deterministic search techniques. However, when dealing with even regular high dimensional problems, it gets difficult or too time consuming for all the model parameters or variables to converge within a given margin of error. In particular, as the dimensions or the number of model parameters increases, so does the required population size due to the exponential scale-up of the search space size. Large population sizes imply large number of cost function evaluations.

The research issue addressed here is that of ensuring constructive population diversity to achieve desirable EA performance for high dimensional optimization problems (Wang et al., 2011). The POPULATION_EA framework proposed in this research is aimed specifically at such problems. While the proposed method is not necessarily scalable to big data optimization problems in its current form, it is a move in the right direction.

1.1. Big data challenges relevant to evolutionary algorithms

Twenty first century has seen the explosion of data collection in many areas. The ease of acquisition of data as a result of technological revolution has made this feasible (Fan, 2013). Massive data and high dimensionality characterizes many contemporary problem domains such as biomedical sciences, engineering, finance and social sciences. This means machine learning problems handling such spatio-temporal problems may have to deal with tens of thousands of features extracted from documents, images and other objects. Typical key features of big data include very large samples as well as very high dimensionality. Sparseness

* Corresponding author.
E-mail addresses: mbhattacharya@csu.edu.au (M. Bhattacharya), mislam@csu.edu.au (R. Islam), jemal.abawajy@deakin.edu.au (J. Abawajy).

http://dx.doi.org/10.1016/j.jnca.2014.07.032
1084-8045/

of data is another feature associated with big data. Big data are often collected over different locations or platforms contributing to measurement errors and experimental variations.

In evolutionary optimization, high dimensionality contributes to deteriorated performance, increased computational cost, algorithmic instability, spurious correlations, and also incidental noise accumulations. The next section discusses the issues related to high dimensionally and evolutionary algorithms, existing remedial approaches as well as their shortcomings.

1.2. High dimensional problem and EA

Some of the reasons why traditional evolutionary algorithms are not adequate for high dimensional problems are associated with the implicit assumptions about evolutionary algorithm.

First, the average absolute fitness of the population increases continually. The population then seems to gradually loose diversity, until the search converges to the global optimum or premature convergence to some false optimum in the search space.

Many research associate the loss of exploratory capability with the loss of diversity and try to increase the diversity of the population to prevent this (Bhattacharya, 2004; Črepinšek et al., 2013; Hu et al., 2005). However, introduction of diversity per se is not sufficient. The more direct reason that EAs may converge to suboptimal solutions is that with increasing average fitness of the population, only the individuals with competitively high fitness are tend to be allowed to survive. New “explorer” individuals sampling fairly different regions in the search space normally have low fitness, until sufficient local exploration i.e. exploitation reveals their beneficial characteristics (Črepinšek et al., 2013). Even the individuals with competitive fitness, because of their sparse distribution in distant regions may disappear as a result of insufficient sampling or genetic drift. This problem is order of magnitude higher in case of high dimensional problems due to the sheer volume of the search space.

Second, EA's response to the characteristics of the search space where problems with enormous (due to high dimensionality) and highly multimodal search spaces are involved.

Many difficult search problems (Bhattacharya, 2007a,b; Bhattacharya, 2008) are made difficult because their search spaces are enormous (due to high dimensionality) and highly multimodal. Many real-world problems fall in this category; for example, even regular engineering design problems may have over 100 design parameters. Here, we have included multimodality along with high dimensionality as in many real world problems, these properties may not occur in isolation.

1.2.1. Enhanced EAs and high dimensional problems

High dimensionality introduces complexities in the search space rendering most common evolutionary algorithms ineffective. There are evidences that even unimodal functions such as the generalized Rosenbrock may show multimodal tendencies in very high dimensional cases. In case of multimodal functions, the number of local optima may increase exponentially with increase in the number of dimensions. The search space and the number of local optima may become too large for EAs to realistically use a population that is large enough to be representative of the actual search space (see Fig. 1).

So far scalability of an evolutionary algorithm is concerned the most crucial task is how deal with the complex search space resulted from high dimensionality. A number of enhanced EAs have been introduced that attempts at manipulating the search space, not necessarily aiming at scalability issues though. Some of the major enhanced or improved variants on traditional EA (as our proposed solution is based on genetic algorithm (GA) we will cite

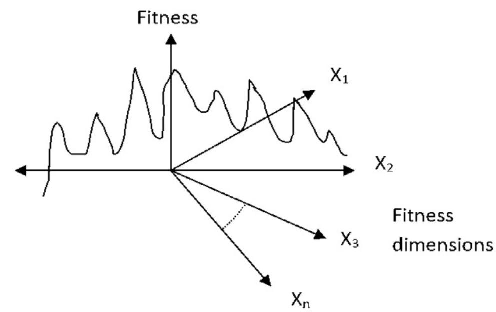

Fig. 1. Relationship between fitness and high dimensionality in a typical EA.

examples of GA variants mainly) are as follows. ***(1) Fitness sharing and crowding:*** Crowding (seminal work by De Jong (1975)) and deterministic crowding (Yannibelli and Amandi, 2012; Mahfoud, 1992) avoids redundancy by replacing similar individuals in the population by different individuals. Bycontrolling the number of individuals in localized search spaces or niches, (Yannibelli and Amandi, 2012) attempted to control the average fitness of the population. Genotypic or phenotypic diversity is maintained in fitness sharing by preventing individuals to gather around high-fitness peaks. Such techniques may be influenced by the properties of the search space (Hu et. al., 2005). ***(2) Multi-population and search space division:*** The multiple subpopulation approach often aims at exploiting each subpopulation by controlling migration. The search space division approach used by Tsutsui et al. (1998) also aimed at working on genetically different populations. In general, both multipopulation and search space division methods focused on population diversity (Yao et al., 2010; Ruciński et. al., 2010). Another diversity-focussed approach was SCGA (Li et al., 2002) which used similarity-based species protection. ***(3) Special learning-based approachs:*** Using a Differential Evolution (DE) implementation, (Wang et. al., 2011) addressed scalability by an enhanced learning-based approach. ***(4) Cooperative coevolution:*** Cooperative coevolution (Li and Yao, 2009; Yang et al., 2008) is perhaps one of the most popular approaches specifically addressing the issue of high dimensionality in EAs. We shall discuss this approach in more details in the next subsection. ***(5) Other approaches:*** Jiang et al. (2011) applied intelligentniche management. An Artificial Bee Colony (ABC) implementation in (Akay and Karaboga, 2012) applied a constraint handling technique into the selection step of ABC to favour the feasible regions of the search space. Lozano et al. (2011) presented some recent works on scalabilityof EAs and other related metaheuristics for continuous optimization problems.

It is common knowledge that EAs do not usually fair very well in terms of scalability and unfortunately majority of the techniques discussed above may require unrealistically large population or large number of subpopulations to be effective while dealing with the “curse of dimensionality”. Historically, it may be noted that Koza et al. (1999) used population sizes of 350,000 or larger in order to attain satisfactory results. However, impractically large population size naturally comes with impractically high computational cost and also increasing the population size does not necessarily improve the performance of an EA. Hence, any technique requiring large population size or large number of subpopulations does not provide realistic solution for scalability in case of high dimensional problems.

1.2.2. High dimensionality specific EAs

One of the early attempts at handling high dimensionality is was by Krishnakumar (1989). He proposed a micro-genetic algorithm that uses micro populations to locate promising areas of the