Expert Systems With Applications 78 (2017) 32–50

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Optimization based on phylogram analysis

Antonio Soares[a,*], Ricardo Râbelo[a], Alexandre Delbem[b]

[a] Federal University of Piauí, Piauí, Brazil
[b] University of São Paulo, Piauí, Brazil

ARTICLE INFO

Article history:
Received 2 September 2016
Revised 3 February 2017
Accepted 4 February 2017
Available online 5 February 2017

Keywords:
Estimation of distribution algorithms
Phylogram analysis
Hierarchical clustering
Combinatorial optimization

ABSTRACT

Estimation of Distribution Algorithms (EDAs) is evolutionary algorithms with relevant performance in handling complex problems. Nevertheless, their efficiency and effectiveness directly depends on how accurate the deployed probabilistic models are, which in turn depend on methods of model building. Although the best models found in the literature are often built by computationally complex methods, whose corresponding EDAs require high running time, these methods may evaluate a lesser number of points in the search space. In order to find a better trade-off between running time (efficiency) and the number of evaluated points (effectiveness), this work uses probabilistic models built by algorithms of phylogenetic reconstruction, since some of them are able to efficiently produce accurate models. Then, an EDA, namely, Optimization based on Phylogram Analysis, and a new search technique, namely, Composed Exhaustive Search, are developed and proposed to find solutions for combinatorial optimization problems with different levels of difficulty. Experimental results show that the proposed new EDA features an interesting trade-off between running time and number of evaluated points, attaining solutions near to the best results found in the literature for each one of such performance measures.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

An important field of research in Evolutionary Algorithms (EAs) consists in developing EDAs (Larrañaga & Lozano, 2001; Pelikan, Goldberg, & Lobo, 2002; Pelikan, Sastry, & Cantú-Paz, 2006) to build probabilistic models of variable values by means of a set of samples selected in promising regions of the search space. From these models, an EDA generates new solutions with values for each variable that are consistent with such regions. In general, a probabilistic model is built at each generation.

The exactitude of a probabilistic model directly affects the EDA's convergence rate (Goldberg, 2002) and its ability to attain the optimal solution of the problem. In other words, the sole use of a probabilistic model does not guarantee neither efficiency nor effectiveness of an EDA; under certain conditions, it may be similar to ordinary EAs (Harik, Lobo, & Goldberg, 1999; Mühlenbein, 1997). In this context, a typical example is the relationship between Genetic Algorithm (GA) and compact GA (Harik et al., 1999); the latter uses a univariate probabilistic model and has similar performance to GA (Harik et al., 1999) for various problems tested in the literature. One advantage of a univariate probabilistic model is that the construction of such a model requires relatively low running time. On the other hand, the Bayesian Optimization Algorithm (BOA) (Pelikan, Goldberg, & Cantú-Paz, 1999) requires significantly more running time to build multivariate and multimodal probabilistic models, namely, Bayesian Networks (Pelikan et al., 1999). This EDA is able to solve certain problems that usually have poor performance for many EAs.

Thus, an important aspect in evaluating the performance of EDAs is the trade-off between the efficiency of the model building process and the exactitude (effectiveness) of the built probabilistic model (Pelikan et al., 2002). The strategy deployed in this work to reach such a trade-off is based on (exact or approximately) decomposing relatively complex problems into simpler sub-problems. Then, from the obtained models for these sub-problems, probabilistic models for the original problem can be efficiently obtained while maintaining their capability of representing promising regions.

The decomposition of the original problem into sub-problems may be achieved by identifying groups of highly correlated variables, which, in some cases, are called Building Blocks BBs. The theory presented in Goldberg (2002) shows that more efficient and effective EDAs, in general, build probabilistic models capable of correctly mapping more BBs by sampling several optimal instances of them. They also present a better use of information from BB to generate new solutions. Moreover, such EDAs must ensure that the optimal instances of BB prevail in relation to sub-optimal

* Corresponding author.
E-mail addresses: ahelsonms@gmail.com (A. Soares), ricardoalr@ufpi.edu.br (R. Râbelo), acbd@icmc.usp.br (A. Delbem).

http://dx.doi.org/10.1016/j.eswa.2017.02.012
0957-4174/© 2017 Elsevier Ltd. All rights reserved.

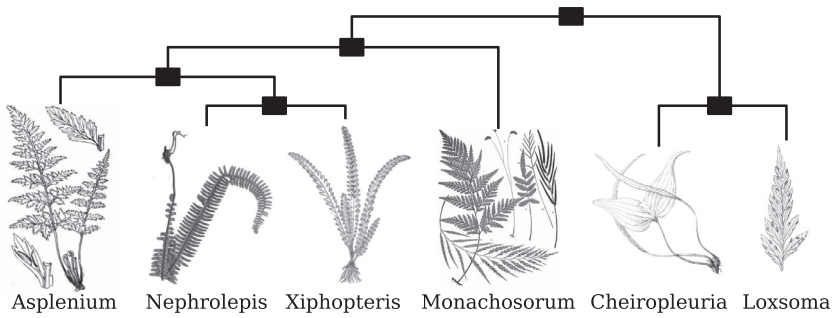

Fig. 1. Phylogeny of a group of plants obtained from test set sequences of nucleotides rbcL_55 (Cancino & Delbem, 2007) and Neighbor Joining algorithm (Saitou & Nei, 1987; Studier & Keppler, 1988).

ones during the evolutionary process and that convergence drift is avoided (Goldberg, 2002).

This work investigates aspects of computational biology methods used to exact or approximately decompose a problem into subproblems. These methods have been successfully used for phylogenetic reconstruction, which provides evolutionary relationships between species and other biological objects (Cancino & Delbem, 2007; Felsenstein, 2003). From such an investigation on phylogenetic methods, we propose a new EDA, namely, Optimization based on Phylogram Analysis (OPA), to attain solutions for combinatorial optimization problems with different levels of difficulty with five main features: (i) the use of a set of trees, instead of only one, to model the correlation between variables; (ii) a summary of relevant information, from the set of trees, to a network structure; (iii) the use of a community detection algorithm (Newman & Girvan, 2004) to identify possible BBs of the problem from one or more network structures; (iv) resampling in several steps of the algorithm to make network structures more representative of correlations between problem variables; and (v) a new search technique, namely, Composed Exhaustive Search (CES).

Features (i) to (iv) are organized into a new method of clustering, namely, Clustering based on Phylogram Analysis (CPA); OPA consists in the association of CPA and CES. Thus, CPA, CES, and OPA, consequently, show how such features can deal with aspects of combinatorial optimization problems that make them complex. The paper presents a theoretical analysis that enables predictions on the performance of OPA; moreover, it shows experimental results that corroborate the theoretical estimates.

The remaining of the paper is organized as follows. Section 2 introduces the concept of phylogeny as a probabilistic model for EDAs. Section 3 describes all steps of CPA. Section 4 shows how possible BBs can be identified for various problems by using CPA. Section 5 describes some EDAs that were also developed to solve combinatorial problems by means of variable partitioning. Section 6 presents CES, which efficiently solves combinatorial optimization problems by means of partitions obtained by CPA. Section 7 shows the results by combining the CPA and CES techniques to solve deceptive trap problems. Section 8 shows the results for OneMax and BinInt problems and with overlapping BBs. Finally, Section 10 summarizes the main contributions of this work.

2. Phylogram-based models

A phylogeny is tree representation of species relationships with common origin. The term phylogenetic tree is used for phylogenies obtained from both morphological data and genetic sequences. In this work, phylogenies from various types of data are built in order to determine groups of objects (phyla), correlating such data.

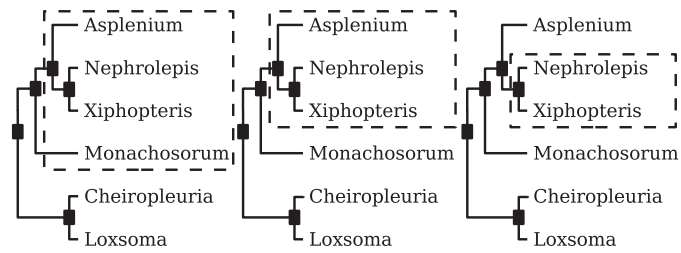

Fig. 2. Possible clades (dashed lines) obtained from the phylogeny in Fig. 1.

Usually, phylogenetic trees (a connected and acyclic graph) are binary trees where their leaves represent species; thus, leaves are labeled with the name of a corresponding species. Fig. 1 shows a phylogeny that highlights evolutionary relationships for a group of plants. The leaves represent existing species, while the internal nodes indicate hypothetical ancestrals or extinct species.

Clades are groups that involve an ancestor and all descendants starting from the lowest (leaf) to a higher (ancestor) internal node in the tree structure. They are, in general, the most useful information in a phylogeny, since relationships within a clade are strong. Fig. 2 illustrates possible clades containing the *Nephrolepis* and *Xiphopteris* species, whose DNA sequences were obtained from test set rbcL_55 (Cancino & Delbem, 2007).

In a computational biology method to solve optimization problems, a clade involves no biological data (except for some bioinformatic problems) and corresponds to a subset of correlated variables, which can be viewed as BBs. A variable can be present in more than one BBs, and BBs can overlap if they have variables in common. In Biology, DNA sequences overlap if they have genetic material in common, as it usually occurs in a clade.

A resampling process (Felsenstein, 2003) can be used to generate various phylograms (phylogenies when DNA sequences are used) from the same data set in order to provide statistic significance to the found relationships. By means of phylograms, CPA can find a consensus of those relationships and compose a network based on them, where highly correlated variables (indicating a BBs) are modeled by a graph clique. Afterwards, the CPA technique uses a community detection algorithm from Complex Networks (Donetti & Muñoz, 2004; Duch & Arenas, 2005) to automatically detect communities (BBs from the optimization point of view) from the network.

3. Description of CPA

CPA can construct models of relationships among variables considering problems with different levels of difficulty, i.e., variables can be partitioned into: (i) BBs with no overlap, (ii) BBs with some overlaps, and (iii) BBs with overlap. CPA has eight main steps that

Download English Version:

https://daneshyari.com/en/article/4943578

Download Persian Version:

https://daneshyari.com/article/4943578

Daneshyari.com