Evolving NK-complexity for Evolutionary Solvers

Roberto Santana
University of the Basque Country (UPV/EHU)
P. Manuel de Lardizabal 1
San Sebastian 20018, Spain
roberto.santana@ehu.es

Alexander Mendiburu
University of the Basque Country (UPV/EHU)
P. Manuel de Lardizabal 1
San Sebastian 20018, Spain
alexander.mendiburu@ehu.es

Jose A. Lozano
University of the Basque Country (UPV/EHU)
P. Manuel de Lardizabal 1
San Sebastian 20018, Spain
ja.lozano@ehu.es

ABSTRACT

In this paper we empirically investigate the structural characteristics that can help to predict the complexity of NK-landscape instances for estimation of distribution algorithms (EDAs). We evolve instances that maximize the EDA complexity in terms of its success rate. Similarly, instances that minimize the algorithm complexity are evolved. We then identify network measures, computed from the structures of the NK-landscape instances, that have a statistically significant difference between the set of easy and hard instances. The features identified are consistently significant for different values of N and K.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*; I.2.8 [**Computing Methodologies**]: Artificial Intelligence—*Problem Solving, Control Methods, and Search*

General Terms

Algorithms

Keywords

estimation of distribution algorithm, NK-model, complexity analysis, networks

1. INTRODUCTION

One of the questions that have traditionally occupied researchers in the evolutionary computation (EC) community is how to characterize, and predict if possible, the difficulty that a given instance poses for an evolutionary algorithm (EA). Different fitness measures have been proposed that quantify a variety of elements related to the behavior of the EAs.

In this paper we propose an approach for the empirical analysis of problem difficulty in instances of the NK-landscape [2] problem that comprises three main steps. Firstly, to evolve the instance structures, keeping the parametrical part intact with the aim to maximize, or minimize, the instance complexity. Secondly, to extract from each evolved instance a detailed characterization in terms of network measures. Finally, use statistical analysis to identify which instance features have a different distribution between the set of easy and hard instances. We apply this procedure to an original set of 9000 instances of the NK-landscape model that were proposed and studied in [3]. The statistical analysis detected that frequencies associated to two network motifs were consistently and significantly different between easy and hard instances across different values of N and K.

GECCO'12 Companion, July 7–11, 2012, Philadelphia, PA, USA.
ACM 978-1-4503-1178-6/12/07.

2. EVOLVING COMPLEXITY

We will measure the instance complexity in terms of the EDA's success rate at solving it. The Estimation of Bayesian networks algorithm (EBNA) [1] enhanced by a local optimization algorithm described in [3] is used for this purpose. For a given NK-landscape instance we assume that the optimum value is known and run EBNA 100 times to determine how many times the optimum is found. This number of times is the fitness $f(G)$ associated to instance G.

To find the optimal instances, we used a random hill climbing algorithm (RHC) that starts from a random instance and randomly modifies its neighborhood structure. The numerical values describing the function potentials are not modified. If the optimization function is improved, then the new instance is accepted, otherwise another possible modification of the neighborhood structure is proposed. The maximum number of evaluations allowed to RHC was 50.

To generate our benchmark, we used an initial dataset of 9000 instances, 1000 for every possible combination of $n \in \{20, 28, 34\}$ and $K \in \{4, 5, 6\}$. All instances were solved using a branch-and-bound algorithm as described in [3]. The branch-and-bound algorithm[1] guarantees that the optimal solutions are found. Every time a new instance is generated using RHC we need to run the branch-and-bound algorithm to compute the new optimum of the NK fitness landscape function. Starting from each of the $9,000$ instances we generated two additional instances, one easy and one hard instance. The final benchmark comprises these additional $18,000$ instances.

3. EXPERIMENTS

The objective of the experiments is to investigate whether the network measures extracted from the evolved instances capture the differences between the sets of easy and hard instances. First, we compute for each instance a large set

[1]We use the implementation by the author, available from `http://medal.cs.umsl.edu/software.php`

of network measures that serve as topological descriptors. Then, we apply a statistical test to each of the features to identify those that have a significantly different distribution between easy and hard instances. Table 1 describes the topological measures extracted from the NK-landscape structures. The computation of the number of structural and functional motifs was implemented using the brain connectivity toolbox [4].

Id	Property	Feat. Number
1	degree	N
2	indegree	N
3	outdegree	N
4	density und.	1
5	density dir.	1
6	assortativity und.	1
7	assortativity dir.	1
8	betweenness	N
9	mean reachability	N
10	mean distance	N
11	characteristic path length	1
12	eccentricity	N
13	radius	1
14	diameter	1
15	clustering coefficient	N
16	shortcuts prob.	$N \cdot (N-1)$
17	range vertex	$N \cdot (N-1)$
18	mean edge range	1
19	fraction shortcuts	1
20	mean motif number $Z=3$	13
21	vertex motif number $Z=3$	$13N$
22	Newmann modularity	1
23	node part. coefficient	N

Table 1: Topological measures extracted from the NK-landscape structures.

In order to identify the set of significant features, we applied, for each feature, a statistical test to determine whether there exists significant difference between the easy and hard instances for the given feature. The statistical test of choice was the Wilcoxon rank sum test of equal medians and the parameter $\alpha = 0.05$ was fixed for all the statistical tests. The test outputs the p-value corresponding to the statistics and we use these values to further characterize the differences between the features.

Table 2 shows which of the network measures described in Table 1 were identified as significant for any combination of N and K. When groups of features were considered, the table shows how many of the features in the group were detected as significant. It can be seen in Table 2 that out of all possible statistical tests only statistical differences between the sets of easy and hard instances are found only 26 times for 8 network measures. There are two coincidences for the clustering coefficient and the node participation coefficient respectively but in every case, only one test of N (there is one coefficient for every node in the network) found significant differences. Therefore 2 tests out of $9(20+28+34) = 738$ tests might be due to multiple testing.

4. CONCLUSIONS

In this paper we have introduced an empirical method for investigating some factors that could predict differences in the complexity of NK-landscape instances for EAs. Our method is based on the direct evolution of easy and hard instances using the success rate of an evolutionary algorithm to estimate the instance complexity. The evolutionary process guarantees that the evolved instance will be "easier" or "harder" than the initial instance. By applying evolution in the two directions of difficulty we can guarantee that the two final sets will differ in terms of complexity with respect to the original set, and more significantly, between them. The type of network measures extracted from the instances can be also applied to the structures of the graphical models for unveiling relevant information from the problem and allowing transfer learning between problem instances [5].

N	20			28			34		
K	4	5	6	4	5	6	4	5	6
9							1		
11			1						
15		1						1	
16			1						
20	2	2	2	2	2	2	2	1	2
21					1				
22		1							
23		1				1			

Table 2: Relevant features identified by the application of the statistical test.

Acknowledgments

This work has been partially supported by the Saiotek and Research Groups 2007-2012 (IT-242-07) programs (Basque Government), TIN2010-14931 and Consolider Ingenio 2010 - CSD 2007 - 00018 projects (Spanish Ministry of Science and Innovation) and COMBIOMED network in computational biomedicine (Carlos III Health Institute)

5. REFERENCES

[1] R. Etxeberria and P. Larrañaga. Global optimization using Bayesian networks. In A. Ochoa, M. R. Soto, and R. Santana, editors, *Proceedings of the Second Symposium on Artificial Intelligence (CIMAF-99)*, pages 151–173, 1999.

[2] S. Kauffman. *Origins of Order*. Oxford University Press, 1993.

[3] M. Pelikan. Analysis of estimation of distribution algorithms and genetic algorithms on NK landscapes. In *Proceedings of the Genetic and Evolutionary Computation Conference GECCO-2009*, pages 1033–1040. ACM, 2008.

[4] M. Rubinov, S. A. Knock, C. J. Stam, S. Micheloyannis, A. W. F. Harris, L. M. Williams, and M. Breakspear. Complex network measures of brain connectivity: Uses and interpretations. *Neuroimage*, 52(3):1059–1069, 2010.

[5] R. Santana, C. Bielza, and P. Larrañaga. Network measures for re-using problem information in EDAs. Technical Report UPM-FI/DIA/2010-3, Department of Artificial Intelligence, Faculty of Informatics, Technical University of Madrid, June 2010.