A New DSM Clustering Algorithm for Linkage Groups Identification

Amin Nikanjam, Hadi Sharifi, B. Hoda Helmi, Adel Rahmani
SCOMAS Lab., Computer Engineering Department, Iran University of Science and Technology
Narmak, Tehran, 16846_13114, Iran
{nikanjam,helmi,rahmani}@iust.ac.ir, hsharifi@comp.iust.ac.ir

ABSTRACT
Linkage learning has been considered as an influential factor in success of genetic and evolutionary algorithms for solving difficult optimization problems. In this paper, a deterministic model named Dependency Structure Matrix (DSM) is used for explicitly decomposing the problem. DSM captures pair-wise dependencies of the problem that must be turned into higher order interactions while solving complex problems. One way to obtain these higher order interactions (linkage groups) is clustering the DSM. A new DSM clustering algorithm is proposed in this paper which is able to identify all the linkage groups from a less accurate DSM leading to a reduction in the number of fitness calls required for identifying the linkage groups. The proposed technique is tested on several benchmark problems and it is shown that it can accurately identify all the linkage groups by $O(n^{1.7})$ fitness evaluations, where n is problem size.

Categories and Subject Descriptors
G.1.6 [**Optimization**]: Global Optimization

General Terms
Algorithms, Performance, Theory.

Keywords
Linkage learning, Dependency structure matrix, DSM clustering.

1. INTRODUCTION
Genetic and evolutionary algorithms have been widely applied to a broad range of optimization problems. For small and simple problems, these algorithms work efficiently and effectively, but for complex problems they are incapable of effectively solving them. The concept of linkage learning arose as a solution to strengthen these algorithms. Linkage learning is finding the relationship between variables. The variables which are related to each other form a linkage group. Once correct linkage groups are found, the problem can be solved easily by a simple GA search. The proposed approach in this paper like the offline utilization of DSMGA [5] uses the DSM, which is calculated based on non-linearity detection of fitness changes [1]. Then a new effective clustering method is applied to identify the linkage groups from pair-wise dependencies in the DSM. In the proposed method, detection of all the linkage groups is done separately before the evolutionary process.

GECCO'10, July 7–11, 2010, Portland, Oregon, USA.
ACM 978-1-4503-0072-8/10/07.

2. DSM BACKGROUND
A DSM by definition is an adjacency matrix of a graph where each entry d_{ij} represents the degree of dependency between node i and node j where each node stands for an element or variable in a complex system [3]. The larger the d_{ij} is, the higher the dependency is between node i and node j. DSM entries d_{ij} can be real numbers or integers. In this paper we focus on the binary domain, where d_{ij}=0 means that there is no dependency between node i and node j, and d_{ij}=1 means that node i and node j are dependent. DSM construction involves detecting dependencies between the problem variables. Several methods are introduced in the literature for calculating pair-wise dependencies and constructing DSM [4]. DSM construction in this paper is based on LINC [1] and is similar to [4, 5]. This method is based on investigating non-linearity of fitness changes by injecting perturbations in a pair of variables. The goal of DSM clustering is to find particular subsets of DSM elements (i.e., clusters) such that variables within a cluster are maximally interacting and variables within different clusters are minimally interacting. DSM clustering is a complicated task which requires expertise [3]. Different methods were introduced in the literature to extract clusters from a DSM [3, 6] that suffer from different problems like oversimplified objective function (predicting good clustering) and restricting parameters like maximum number of clusters.

3. THE PROPOSED DSM CLUSTERING ALGORITHM
In this paper, a new effective DSM clustering method is introduced that does not need any objective function or metric. The proposed clustering strategy is based on natural groups of variables that were seen in the DSM and comprised of several phases to generate and revise these initial groups. The first phase starts with grouping the variables (namely initial clusters) which are dependent to each other based on density of groups (Equation 1) according to the DSM constructed before (Algorithm 1). The second phase involves constructing secondary clusters using initial clusters. This is done using moving variables between clusters based on co-occurrence of variables in the clusters. The third phase of clustering consists of revising secondary clusters based on DSM entries to obtain the final clusters [2]. Theoretical analysis reveals that the number of fitness calls needed to detect all the pair-wise non-linearity is bounded by $O(n^2)$ but the experimental results presented in the next section show that it is $O(n^{1.7})$ for the proposed method to extract all the linkage groups. Time complexity of the proposed clustering algorithm is shown to be $O(n^3)$ [2].

Algorithm 1: DSM Clustering, Phase 1

for each variable i in the variable set,
  create a new group, gp_i containing variable i.
  for each variable j,
    if ($DSM[i,j]$==1),
      add j to gp_i.
for each gp_i,
  while((Density(gp_i)<0.5) or (there is a weak variable that dependant to less than the half of variables in gp_i))
    find the weakest variable and remove it from gp_i
  initCl= gp_i

$$Density(G) = \frac{\sum_{i,j \in G} DSM[i,j]}{|G|^2} \quad (1)$$

where G is a group of variables in the DSM and |G| is the number of variables in G.

4. EXPERIMENTAL RESULTS

The proposed linkage identification algorithm is evaluated by two performance measures: accuracy of DSM and the number of fitness evaluations needed to identify all the linkage groups. Fitness evaluations are needed for the DSM construction phase and there is no need to evaluate fitness function during DSM clustering. Table 1 shows a comparison between the proposed approach and the only comparable method to our work, offline utility of DSMGA [5]. DSM accuracy is defined as the ratio of the number of fully represented linkage groups in the DSM to the number of linkage groups that should have been represented in the DSM if it has been constructed flawlessly. The DSM accuracy will be 1 if all the correct pair-wise dependencies are captured in DSM. Experimental results in this paper and another work [5] reveal that as the number of fitness function evaluated to construct the DSM increases, ratio of correct extracted linkage groups grows. Results show the effectiveness of the proposed clustering algorithm in extracting all the linkage groups from a less accurate DSM. Figure 1 (loglog scale) illustrates the number of fitness evaluations needed to correctly identify all the linkage groups by the proposed DSM clustering approach applied to different trap functions [2]. All the experimental results are averaged over 30 independent runs. Empirical results show scalability of the proposed algorithm with respect to the number of fitness evaluations.

Table 1. Comparison between the proposed method and offline utility of DSMGA [5] on a 50 bits Trap5 function

Method	Number of fitness evaluations	DSM Accuracy	Proportion of correct linkage groups
The proposed method	7000	0.723	1
Offline DSMGA	11712	0.906	0.998

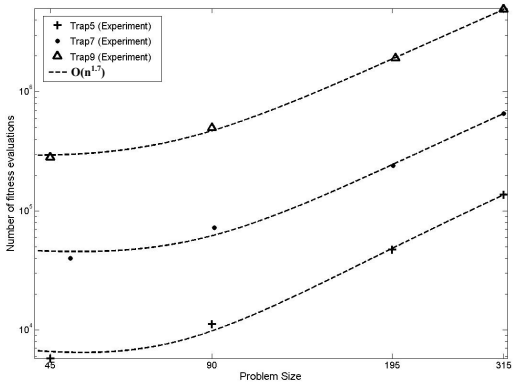

Figure 1. Number of fitness evaluations needed for identifying all linkage groups in different trap functions.

5. CONCLUSIONS

In this paper, a new DSM clustering algorithm has been presented for linkage group identification. DSM entries that represent pair-wise dependencies were calculated by inspecting the fitness changes caused by perturbation. DSM can capture all the dependencies if the number of individuals evaluated to construct the DSM is large enough. A new clustering algorithm was proposed to extract all the linkage groups from the constructed DSM. The proposed DSM clustering algorithm was shown to be able to identify all the linkage groups from a less accurate DSM. This leads to a reduction in the number of fitness evaluations required for extracting all the linkage groups from the theoretical bound $O(n^2)$ to $O(n^{1.7})$. The proposed DSM clustering algorithm does not impose any additional load in terms of fitness evaluations. Experimental results have revealed the scalability of the algorithm in problems with different sizes and different orders of difficulty.

6. REFERENCES

[1] Munetomo, M. and Goldberg, D.E. 1999. Identifying linkage groups by nonlinearity/non-monotonicity detection. In Proceedings of the 1999 Genetic and Evolutionary Computation Conference (Orlando, USA).

[2] Nikanjam, A., Sharifi, H., Helmi, B.H. and Rahmani, A. 2010. Enhancing the Efficiency of Genetic Algorithm by Identifying Linkage Groups Using DSM Clustering. In Proceedings of 2010 IEEE Congress on Evolutionary Computation (CEC2010) (Barcelona, Spain).

[3] Sharman, D.M. and Yassine, A.A. 2004. Characterizing complex product architectures. Systems Engineering, 7 (1), 35-60.

[4] Yu, T.L. 2006. A matrix approach for finding extreme: Problems with modularity, hierarchy, and overlap. Doctoral Thesis. University of Illinois at Urbana-Champaign.

[5] Yu, T.L. and Goldberg, D.E. 2004. Dependency structure matrix analysis: Offline utility of the dependency structure matrix genetic algorithm. LNCS 3103. Springer. 355-366.

[6] Yu, T.L., Goldberg, D.E., Sastry, K., Lima, C.F. and Pelikan, M. 2009. Dependency structure matrix, genetic algorithms, and effective recombination. Evolutionary Computation, 17 (4). 595-626.