MB-GNG: Addressing Drawbacks in Multi-Objective Optimization Estimation of Distribution Algorithms

Luis Martí[a,*], Jesús García[a], Antonio Berlanga[a], Carlos A. Coello Coello[b], José M. Molina[a]

[a] *Group of Applied Artificial Intelligence, Department of Informatics, Universidad Carlos III de Madrid. Av. de la Universidad Carlos III, 22. Colmenarejo 28270 Madrid, Spain.*
[b] *Department of Computer Science, CINVESTAV-IPN, Av. IPN No. 2508, Col. San Pedro Zacatenco. México, D.F. 07360, Mexico.*

Abstract

In this paper we examine the model-building issue related to multi-objective estimation of distribution algorithms (MOEDAs) and show that some of their, as yet overlooked, characteristics render most current MOEDAs unviable when addressing optimization problems with many objectives. We propose a novel model-building growing neural gas (MB-GNG) network that is specially devised for properly dealing with that issue and therefore yields a better performance. A series of experiments are conducted in order to show from an empirical point of view the advantages of the new algorithm.

Keywords: Multi-objective optimization, estimation of distribution algorithm, model building, growing neural gas

1. Introduction

Most human endeavors involve the creation of artifacts with properties that must be tuned to be as efficient as possible. This fact has prompted the creation of a number of interrelated research areas like optimization, mathematical programming, operational research and decision-making. Although these areas share some of their goals, each of them differs from the others on the approaches put forward by their respective communities and the characteristics of the problems they deal with.

Many real-world optimization problems involve more than one goal to be optimized. This type of problems is known as *multi-objective optimization problems* (MOPs). A MOP can be expressed as the problem in which a set of *objective functions*

*Corresponding author.
Email addresses: `lmarti@inf.uc3m.es` (Luis Martí), `jgherrer@inf.uc3m.es` (Jesús García), `aberlan@ia.uc3m.es` (Antonio Berlanga
), `ccoello@cs.cinvestav.mx` (Carlos A. Coello Coello), `molina@ia.uc3m.es` (José M. Molina)
URL: `http://www.giaa.inf.uc3m.es/miembros/lmarti` (Luis Martí)

marti–mb-gng–or-letters.tex 1248 2010-11-24 04:24:46Z lm-deep

$f_1(\boldsymbol{x}), \ldots, f_M(\boldsymbol{x})$ should be jointly optimized;

$$\min \ \boldsymbol{F}(\boldsymbol{x}) = \langle f_1(\boldsymbol{x}), \ldots, f_M(\boldsymbol{x}) \rangle \, ; \ \boldsymbol{x} \in \mathcal{S} \, ; \tag{1}$$

where $\mathcal{S} \subseteq \mathbb{R}^n$ is known as the *feasible set* and could be expressed as a set of restrictions over the decision set, $\mathbb{R}^n$. The image set of $\mathcal{S}$ produced by function vector $\boldsymbol{F}(\cdot)$, $\mathcal{O} \subseteq \mathbb{R}^M$, is called *feasible objective set* or criterion set (see [1, 2] for details on notation).

In this class of problems the optimizer must find one or more feasible solutions that jointly minimize (or maximize) the objective functions. Therefore, the solution to this type of problem is a set of trade-off points. The optimality of a solution can be expressed in terms of the Pareto dominance relation.

Definition 1 (Pareto dominance relation). For the optimization problem specified in (1) and having $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}$, $\boldsymbol{x}$ is said to dominate $\boldsymbol{y}$ (expressed as $\boldsymbol{x} \prec \boldsymbol{y}$) iff $\forall f_j, \ f_j(\boldsymbol{x}) \leq f_j(\boldsymbol{y})$ and $\exists f_i$ such that $f_i(\boldsymbol{x}) < f_i(\boldsymbol{y})$.

Definition 2 (non-dominated subset). In problem (1) and having the set $\mathcal{A} \subseteq \mathcal{S}$. The *non-dominated subset* of $\mathcal{A}$, $\hat{\mathcal{A}} \subseteq \mathcal{A}$, is defined as

$$\hat{\mathcal{A}} = \left\{ \boldsymbol{x} \in \mathcal{A} \middle| \nexists \boldsymbol{x}' \in \mathcal{A} : \boldsymbol{x}' \prec \boldsymbol{x} \right\} . \tag{2}$$

Definition 3. The solution of (1) is $\hat{\mathcal{S}}$, the non-dominated subset of $\mathcal{S}$.

$\hat{\mathcal{S}}$ is known as the *efficient set* or *Pareto-optimal set* [2]. The elements of this set are to be assessed and evaluated by a *decision maker* in order to select which one(s) will be part of the final solution of the problem. It is worth noting that some alternative notations, like *non-inferior* or *non-dominated set*, can be found in the literature (again, see [1, 2]).

If problem (1) has certain characteristics, e. g., linearity or convexity of the objective functions or convexity of $\mathcal{S}$, the efficient set can be determined by mathematical programming approaches. However, in the general case, finding the solution of (1) is an *NP*–complete problem [3]. In this case, heuristic or metaheuristic methods can be applied in order to have solutions of practical value at an admissible computational cost.

A broad range of heuristic and metaheuristic approaches has been used to address MOPs [4]. Of these, multi-objective evolutionary algorithms (MOEAs) [5] have been found to be a competent approach in a wide variety of application domains. Their main advantages are ease of use, inherent parallel search and lower susceptibility to the shape or continuity of the image of the efficient set, compared with traditional mathematical programming techniques for multi-objective optimization [4].

There is a class of MOPs that are particularly appealing because of their inherent complexity: the so-called many-objective problems [6]. These are problems with a relatively large number of objectives (normally, four or more). Although somewhat counterintuitive and hard to visualize for a human decision maker, these problems are not uncommon in real-life engineering practice. For example, [7] details some relevant real problems of this type.

marti–mb-gng–or-letters.tex 1248 2010-11-24 04:24:46Z lm-deep

The scalability issues of traditional MOEAs in these problems have triggered a sizable amount of research, aiming to provide alternative approaches that can properly handle many-objective problems and perform reasonably.

Estimation of distribution algorithms (EDAs) are one such approaches [8]. EDAs have been hailed as a paradigm shift in evolutionary computation. They build a model of the population instead of applying evolutionary operators. This model is then used to synthesize new individuals. Probably because of their success in single-objective optimization, EDAs have been extended to the multi-objective optimization problem domain, leading to the so-called multi-objective EDAs (MOEDAs) [9].

Although MOEDAs have yielded some encouraging results, their introduction has not lived up to *a priori* expectations. This can be attributed to a number of different causes. We have recognized three of them, in particular, those derived from the incorrect treatment of population outliers; the loss of population diversity, and that too much computational effort is being spent on finding an optimal population model.

A number of works have dealt with the issues listed above, particularly with loss of diversity. Nevertheless, in our opinion, the community has failed to acknowledge that the underlying cause for all those problems could, perhaps, be traced back to the algorithms used for model building in EDAs.

In this paper we examine the model-building issue of current MOEDAs and show that some of its characteristics, which have been disregarded so far, render most current approaches unsuitable for tackling MOPs. We then propose a novel model-building algorithm, based on the growing neural gas (GNG) network. This model-building GNG (MB-GNG) is the main contribution of this paper. It has been devised with this particular problem in mind, and therefore addresses the problems of current approaches.

The remainder of this paper is organized as follows. Section 2 serves as a brief introduction to MOEDAs and the issues present in current model-building algorithms. After this, MB-GNG is described in Section 4. Then, in Section 5, a comparative study is carried out in order to establish from an experimental point of view the improvements introduced by MB-GNG with respect to similar algorithms. Finally, some conclusive remarks are put forward.

2. Multi-Objective Estimation of Distribution Algorithms

Estimation of distribution algorithms (EDAs) are population-based optimization algorithms. Instead of applying evolutionary operators to the population like other evolutionary approaches, EDAs build a statistical model of the most promising subset of the population. The introduction of machine learning techniques implies that these new algorithms lose the straightforward biological inspiration of their predecessors. Nonetheless, they gain the capacity of scalably solving many challenging problems, in some cases significantly outperforming standard EAs and other optimization techniques.

Multi-objective EDAs (MOEDAs) [9] are the extensions of EDAs to the multi-objective domain. Most MOEDAs consist of a modification of existing EDAs whose fitness assignment function is substituted by one taken from an existing MOEA.

In general terms, MOEDAs follow a common algorithmic scheme (see Figure 1). At a given iteration t, a MOEDA has a population $\mathcal{P}_t$ of individuals, each one repre-

marti–mb-gng–or-letters.tex 1248 2010-11-24 04:24:46Z lm-deep

Parameters:

▹ $n_{\text{pop}} \in \mathbb{N}$, population size.

▹ $\alpha \in (0, 1]$, selection percentile.

▹ $\omega \in (0, 1]$, substitution percentile.

▹ stop : $\mathfrak{P} \to \{\text{true}, \text{false}\}$, stopping criterion function over $\mathfrak{P}$, the ordered collection of populations.[a]

Algorithm:

$t \leftarrow 0$.

Generate the initial population $\mathcal{P}_0$ with n_{pop} individuals.

$\mathfrak{P} = \{\mathcal{P}_0\}$.

repeat

Rank individuals in $\mathcal{P}_t$ according to the fitness assignment function.

Extract the subset $\mathcal{M}_t$ with the best $\lfloor \alpha\, n_{\text{pop}} \rfloor$ elements of $\mathcal{P}_t$.

Build model of $\mathcal{M}_t$.

Sample $\lfloor \omega\, n_{\text{pop}} \rfloor$ new individuals from the model to create $\mathcal{P}'_t$.

Substitute part of $\mathcal{P}_t$ with $\mathcal{P}'_t$ to create $\mathcal{P}_{t+1}$.

$t = t + 1$.

$\mathfrak{P} = \mathfrak{P} \cup \{\mathcal{P}_t\}$.

until stop($\mathfrak{P}$) = true.

return $\hat{\mathcal{P}}$, the set of non-dominated individuals of $\mathcal{P}_t$.

[a] Stopping a multi-objective optimization is a non-trivial task. We refer the interested reader to [10, 11]

Figure 1: A general MOEDA algorithm.

senting a point in the search space. In every iteration, $\mathcal{P}_t$ elements are ranked according to a given fitness assignment function. A subset $\mathcal{M}_t$, with the best elements of $\mathcal{P}_t$ is computed. The model-building algorithm relies on $\mathcal{M}_t$ to create a model of the best part of the population. This model is sampled in order to create new elements which are combined with $\mathcal{P}_t$ to create the population to be used in the next iteration, $\mathcal{P}_{t+1}$.

A given stopping criterion determines when the optimization process should be interrupted. When this happens, $\hat{\mathcal{P}}$, the set of non-dominated individuals of $\mathcal{P}_t$, is returned as solution.

Although there are different approaches for determining $\mathcal{M}_t$ and $\mathcal{P}_{t+1}$, MOEDAs are better characterized by their two main components, the fitness assignment function and the model-building algorithm.

Fitness functions have been mostly taken from MOEAs and can be classified in three main groups: objective function aggregation, Pareto dominance-based ranking, and indicator-based ranking. An in-depth discussion of these alternatives is out of the scope on this paper, but it should be noted that the Pareto dominance-based approach proposed by the NSGA-II algorithm is, by far, the most popular in the current literature. See [5] for a survey on this topic.

The model-building algorithm is the kernel of an EDA. There are two main types of methods for addressing this problem: those based on graphical models and those

marti–mb-gng–or-letters.tex 1248 2010-11-24 04:24:46Z lm-deep

based on mixture distributions.

2.1. Graphical Model MOEDAs

Most EDAs based on graphical models rely on Bayesian networks. From these, the Bayesian optimization algorithm (BOA) [12] is the specific approach that has been extrapolated to the multi-objective domain.

The exhaustive synthesis of a Bayesian network from the algorithm's population is an NP-hard problem [13, 14]. Therefore, these EDAs must employ heuristic alternatives for building their networks while keeping the computational cost under reasonable margins.

BOA-based MOEDAs combine the Bayesian model-building scheme with an already existing Pareto-based fitness assignment. This is the case of the multi-objective BOA (mBOA) [15] that exploits the fitness assignment used in NSGA-II. Another algorithm based on hierarchical BOA (hBOA), called mhBOA [16], also uses the same form of fitness assignment but introduces clustering in the feasible objective set. A similar idea is proposed in [17], where the mixed BOA (mBOA) is combined with SPEA2's selection scheme [5] to form the multi-objective mBOA (mmBOA). The multi-objective real BOA (MrBOA) [18] also extends a preexisting EDA, namely, the real BOA (rBOA). MrBOA combines the fitness assignment of NSGA-II with rBOA.

2.2. Mixture Distribution MOEDAs

Another approach to modelling the subset with the best population elements is to apply a distribution mixture approach. Bosman and Thierens [19] proposed several variants of their multi-objective mixture-based iterated density estimation algorithm (MIDEA). They are based on their IDEA framework. They also introduced a novel Pareto-based and diversity-preserving fitness assignment function. The model construction is inherited from the single-objective version. The proposed MIDEAs considered several types of probabilistic models for both discrete and continuous problems.

MIDEAs do not provide a specific mechanism to ensure equal coverage of the Pareto-optimal front if the number of representatives in some parts of the front is much larger than the number of representatives in some other parts. The clustering algorithms applied for this task include the randomized leader algorithm [20], the k-means algorithm and the expectation-maximization algorithm [21].

MIDEAs are not the only mixture-based algorithms. The multi-objective Parzen EDA (MOPED) [22] puts forward a similar mixture-based approach. MOPED uses the NSGA-II ranking method and the Parzen estimator.

2.3. Other MOEDA Approaches

There are some other approaches for model building. Some take advantage of the mathematical properties of the Pareto-optimal front. For example, the regularity model-based multi-objective estimation of distribution algorithm (RM-MEDA) [23] is based on the regularity property derived from the Karush–Kuhn–Tucker condition. At each iteration, RM-MEDA models the most promising area of the decision space using a probability distribution whose centroid is a $(M - 1)$-dimensional piecewise continuous manifold. Again, this model adopts the fitness assignment of NSGA-II.

marti–mb-gng–or-letters.tex 1248 2010-11-24 04:24:46Z lm-deep

Its main drawback is its high computational requirements, that render it unsuitable for problems with many objective functions.

The covariance matrix adaptation evolution strategies (CMA-ES) [24] have been used in the multi-objective context. CMA-ES consists of a method for updating the covariance matrix of the multivariate normal mutation distribution used in an evolution strategy. They can be viewed as EDAs, as new individuals are sampled according to the mutation distribution. Adaptation of the covariance matrix is equivalent to learning a second-order model of the underlying objective function.

3. Model-Building in the Multi-Objective Case

Regardless of the many efforts at providing usable model-building methods for EDAs, the nature of the problem itself has received relatively little attention. In spite of the succession of gradually improving results of EDAs, one question hangs over the search for possibilities for further improvement: would current statistically sound and robust approaches be valid for the problem being addressed? Generally, model-building algorithms are off-the-shelf machine learning methods that were originally intended for other classes of problems. On the other hand, the model-building problem has particular requirements that the above methods do not meet and may even go against.

In this paper we argue that the model-building problem has not been properly identified. For this reason, it has been treated like other previously existing problems overlooking that this problem has particular requirements. This matter did not show up as clearly in single-objective EDAs. Thanks to the extension to the multi-objective domain this issue has become more evident, as we will debate in the remainder of this section.

There are at least three properties of current model-building approaches that hinder their performance, in particular,

1. the incorrect treatment of data outliers;
2. the loss of population diversity; and
3. the excess of computational effort devoted to finding an optimal population model.

The data outliers' issue is a good example of the defective understanding of the nature of the model-building problem, and is, in our opinion the cornerstone for reaching a better understanding of the problem.

In machine-learning practice, outliers are handled as noisy, inconsistent or irrelevant data. Therefore, outlying data is expected to have little influence on the model or it is just disregarded. However, this behavior is not appropriate for model building. In this case, it is known beforehand that all elements in the data set should be taken into account, as they represent newly discovered or candidate regions of the search space and, therefore, must be explored. Therefore, these instances should be at least equally represented by the model and perhaps even reinforced. A model-building algorithm that properly handles outliers might actually speed up the search process and lower the rate of the exponential dimension-population size dependency.

Another weakness of most MOEDAs (and most EDAs, for that matter) is the loss of population diversity. This is a point that has already been made, and some proposals

marti–mb-gng–or-letters.tex 1248 2010-11-24 04:24:46Z lm-deep

for addressing the issue have been laid out [25, 26, 27]. This loss of diversity can be traced back to the above outliers' issue of model-building algorithms. The repetitive application of an algorithm that disregards outliers tends to generate more individuals in areas of the search space that are more densely represented. Although there have been some proposals to circumvent this problem, we take the view that the ultimate solution is the use of an appropriate algorithm.

The third issue to be dealt with is the computational resources wasted on finding an optimal description for the subpopulation being modelled. In the model-building case, optimal model complexity can be sacrificed in the interests of a faster algorithm. This is because the only constraint is to have a model that is sufficiently, but not necessarily optimal in terms of complexity, in order to represent the data in a correct manner. This is particularly true when dealing with high-dimensional MOPs, as, in these cases, there will be large amounts of data to be repeatedly processed at every iteration. Even so, most current approaches spend considerable effort on determining the optimal model complexity, using minimum description length, structural risk minimization, Bayesian information criterion or other similar heuristics.

In conclusion, we can deduce that understanding the nature of the model-building problem and the application of suitable algorithms seem to point the way forward in this area.

4. Model-Building Growing Neural Gas

Clustering algorithms [21] have been used as part of the model-building algorithms of EDAs and MOEDAs. However, as we discussed in the previous section, a custom-made algorithm might be one of the ways of achieving a significant improvement in this field.

The growing neural gas (GNG) network [28] has been chosen as a starting point after surveying the literature for suitable candidates. GNG networks are intrinsic self-organizing neural networks based on the neural gas [29] model. The term "neural gas" refers to the behaviour of the center of the nodes during the adaptation process, which distribute themselves like a gas within an imaginary container defined by the bounds implicitly given by the dataset on which the network is being trained.

Among the vast number of existing clustering methods we decided to base our approach on GNG because of its interesting properties, in particular:

- the network has been shown to be sensitive to outliers [30], something undesirable in typical applications but suitable for model-building;
- the network adapts its topology automatically to meet the complexity of the problem being solved;
- it has a fast convergence to low distortion errors and these errors are better than those yielded by "standard" algorithms like k-means clustering, maximum-entropy clustering and Kohonen's self-organizing feature maps [29];
- although it benefits from the topological ordering of the nodes it does not suffer the problem associated to Kohonen networks, where a node can pull its neighbors

marti–mb-gng–or-letters.tex 1248 2010-11-24 04:24:46Z lm-deep

to invalid or non-representative locations of the network's input set (which, in our case, is the decision set), and;

- the introduction of a novel cluster repulsion mechanism fosters the exploration of the input space.

A GNG network creates an ordered topology of classes of dataset elements and associates a cumulative error to each. The topology and the cumulative errors are conjointly used to determine how new classes should be inserted. Using these heuristics the model can fit the network's dimension to the complexity of the problem being solved.

Our model-building GNG (MB-GNG) is an extension of the original GNG. It introduces a cluster repulsion term that fosters a better spread of the clusters along the training dataset, as explained in [31].

MB-GNG is a one-layer network that defines each class as a local Gaussian density and adapts them using a local learning rule. The layer contains a set of nodes $\mathcal{C} = \{c_1, \ldots, c_{N^*}\}$, with $N_0 \leq N^* \leq N_{\max}$. Here N_0 and $N_{\max}$ represent initial and maximal number of nodes in the network. The network receives inputs $\boldsymbol{x} \in \mathcal{I}$ in its input set, $\mathcal{I}$. In our case this input set is the decision set: $\mathcal{I} = \mathbb{R}^n$.

A node c_i describes a local multivariate Gaussian density that consists of a center, $\boldsymbol{\mu}_i$, and a standard deviations vector, $\boldsymbol{\sigma}_i$. It also has an accumulated error, ξ_i, and a set of edges that define the set of topological neighbors of c_i, $\mathcal{V}_i$. Each edge has an associated age, ν_{ij}.

The network is initialized with N_0 nodes with their centers set to randomly chosen inputs. A training iteration starts after an input $\boldsymbol{x}$ is randomly selected from the training data set. Then, two nodes are selected for being the closest ones to $\boldsymbol{x}$. The *best-matching node*, c_b,

$$b = \underset{i=1,..,N^*}{\arg\min}\, d\left(\boldsymbol{\mu}_i, \boldsymbol{x}\right), \tag{3}$$

is the closest node to $\boldsymbol{x}$. Consequently, the *second best-matching node*, $c_{b'}$, is determined as

$$b' = \underset{i=1,..,N^*; i\neq b}{\arg\min}\; d\left(\boldsymbol{\mu}_i, \boldsymbol{x}\right). \tag{4}$$

Here $d\left(\boldsymbol{a}, \boldsymbol{b}\right)$ is a metric, in our case, the Euclidean one.

If $c_{b'}$ is not a neighbor of c_b then a new edge $\mathcal{V}_b$ is established between them, where $\mathcal{V}_b = \mathcal{V}_b \cup \{c_{b'}\}$ with zero age, $\nu_{bb'} = 0$. If, on the other hand, $c_{b'} \in \mathcal{V}_b$ the age of the corresponding edge is reset to $\nu_{bb'} = 0$.

At this point, the age of all edges is incremented by one. If an edge is older than the maximum age, $\nu_{ij} > \nu_{\max}$, then the edge is removed. If a node becomes isolated from the rest of the nodes because no edges connecting it remain, it is also deleted.

The clustering error is then added to the best-matching node error accumulator,

$$\Delta\xi_b = d\left(\boldsymbol{\mu}_i, \boldsymbol{x}\right)^2 \ . \tag{5}$$

After that, learning takes place in the best-matching node and its neighbors with rates ϵ_{best} and ϵ_{vic} ($\epsilon_{\text{best}} > \epsilon_{\text{vic}}$), respectively. These two rates gate the movement of the centers of the nodes involved towards the current input $\boldsymbol{x}$.

marti–mb-gng–or-letters.tex 1248 2010-11-24 04:24:46Z lm-deep

For c_b, adaptation follows the rule originally used by GNG,

$$\Delta\boldsymbol{\mu}_b = \epsilon_{\text{best}}\left(\boldsymbol{x} - \boldsymbol{\mu}_b\right) . \tag{6}$$

However, for the neighbors of c_b, a cluster repulsion term [31] is added to the original formulation. This repulsion term avoids the meaningless concentration of nodes in the data space and therefore, promotes a proper representation of the data set with fewer nodes.

Following that, the learning rule for those nodes can be expressed as, $\forall c_v \in \mathcal{V}_b$,

$$\Delta\boldsymbol{\mu}_v = \epsilon_{\text{vic}}\left(\boldsymbol{x} - \boldsymbol{\mu}_v\right) + \beta e^{\left(-\frac{d(\boldsymbol{\mu}_v,\boldsymbol{\mu}_b)}{\zeta}\right)} \frac{\sum_{c_u \in \mathcal{V}_b} d\left(\boldsymbol{\mu}_u, \boldsymbol{\mu}_b\right)}{|\mathcal{V}_b|} \frac{\left(\boldsymbol{\mu}_v - \boldsymbol{\mu}_b\right)}{d\left(\boldsymbol{\mu}_v, \boldsymbol{\mu}_b\right)} . \tag{7}$$

This approach was already used as part of the robust GNG [30] and it has proven itself useful for obtaining a good spread of the clusters in the inputs' space. In the aforementioned work, it was stated that the adaptation rule is not sensitive to its parameters. We have set them to $\beta = 2$ and $\zeta = 0.1$ as suggested in [30].

After a given number, T_+, of dataset iterations have taken place, it can be assumed that there is enough information stored in the error accumulators, ξ_i. This information is used to determine where to add new nodes to the network. In particular, if the current iteration is an integer multiple of T_+ and the network has not reached its maximum size ($N^* < N_{\max}$) then a new node is inserted in the network.

First, the node with the largest error, c_e, is selected. Then, the worst node among its neighbors, $c_{e'}$, is located. Then N^* is incremented and the new node, c_{N^*}, is inserted between the two nodes,

$$\boldsymbol{\mu}_{N^*} = 0.5\left(\boldsymbol{\mu}_e + \boldsymbol{\mu}_{e'}\right); \; \xi_{N^*} = 0.5(\xi_e + \xi_{e'}). \tag{8}$$

The edge between c_e and $c_{e'}$ is removed and two new edges connecting c_{N^*} with c_e and $c_{e'}$ are created. The accumulated errors are decreased

$$\xi_e = \delta_I \xi_e, \; \xi_{e'} = \delta_{\text{I}} \xi_{e'}, \tag{9}$$

by a rate $0 \leq \delta_{\text{I}} \leq 1$. Finally, the errors of all nodes are decreased by a factor δ_{G},

$$\xi_i = \delta_{\text{G}} \xi_i, \; i = 1, .., N^*. \tag{10}$$

The algorithm stops after a learning epoch if the standard deviation of the accumulated errors is smaller than a certain threshold, ρ,

$$\sqrt{\frac{1}{N^*} \sum_{i=1}^{N^*} (\xi_i - \overline{\xi})^2} < \rho. \tag{11}$$

This means that it will stop when the outliers are as well represented as possible.

After training has ended the deviations, $\boldsymbol{\sigma}_i$, of the nodes must be computed. For this task we employ the unbiased normal estimator of the deviations [32].

marti–mb-gng–or-letters.tex 1248 2010-11-24 04:24:46Z lm-deep

The local Gaussian densities resulting from the described algorithm can be combined to synthesize the Gaussian mixture with parameters $\boldsymbol{\Theta}$,

$$P(\boldsymbol{x}|\boldsymbol{\Theta}) = \frac{1}{N^*} \sum_{i=1}^{N^*} P(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i) . \tag{12}$$

Each Gaussian density is formulated as

$$P(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i)\right) , \tag{13}$$

with the covariance matrices $\boldsymbol{\Sigma}_i$ defined as a diagonal matrix with its non-zero elements set to the values of the deviations $\boldsymbol{\sigma}_i$. The Gaussian mixture can be used by the EDA to generate new individuals. These new individuals are created by sampling $P(\boldsymbol{x}|\boldsymbol{\Theta})$. The generation of randomly distributed numbers that follow a given distribution has been dealt in depth by many authors. In our case, we applied the Box–Muller transformation [33].

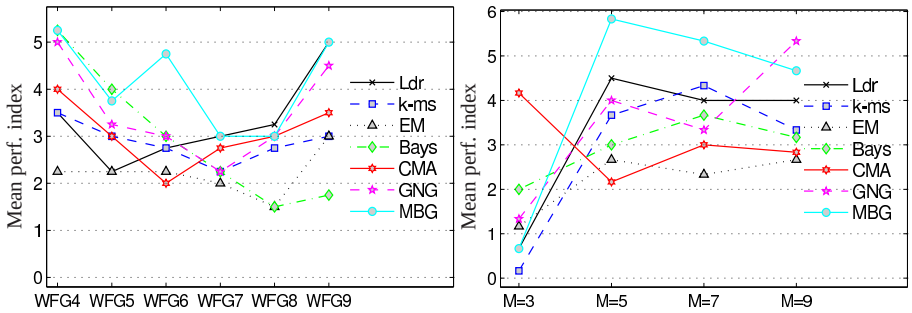
5. Experimental Study

A comprehensive comparative study is essential in order verify that the novel MB-GNG actually yields a substantial improvement. In these experiments we compare MB-GNG and some of the alternative model-building algorithms under a common EDA framework like the one presented in Section 2. The model-building algorithms that also take part in the tests are: Bayesian networks, as used in MrBOA; randomized leader algorithm, k-means algorithm and E-M algorithm, as described for MIDEAs; $(1+\lambda)$-CMA-ES as described in [24]; and the original GNG. Furthermore, the NSGA-II and SPEA2 MOEAs [5] were also involved in the experiments in order to provide a baseline for the comparisons. The performance of each algorithm is assessed in terms of approximation to the Pareto-optimal front and the computational cost of each one.

Six well known community-accepted problems are addressed: the WFG4 to WFG9 problems [34]. These problems pose different classes of challenges, like local-optima, parameter-bias, uni- and multi-modality, etc., to the optimizer. See the corresponding paper for a full description of each one. Each problem is configured with 3, 5, 7 and 9 objective functions. For all cases, the decision space dimension was set at 15. All the algorithms were executed 30 times for each problem/dimension pair.

The quality of the solutions is determined by the use of the hypervolume indicator [35]. This is the only indicator that has the properties of a metric and the only to be strictly Pareto monotonic. However, it must be noted that there are other alternatives.

Statistical hypotheses tests have to be applied to establish the validity of the results. For this, we perform a Kruskal–Wallis test [36] with the indicator values yielded by each algorithm's run for each problem/dimension combination. In the context of these experiments, the null hypothesis for the test was that all algorithms were equally capable of solving the problem. If the null hypothesis was rejected, which was the case in all the experimental instances, the Conover–Inman procedure [36, pp.288–290] was



(a) $\overline{P}_p$ (), mean values for each problem. (b) $\overline{P}_m$ (), mean values for each dimension.

Figure 2: Mean values of the performance index across the different problems and objective space dimensions.

applied in a pairwise manner to determine if the results of one algorithm were significantly better than those of the other. A significance level, α, of 0.05 was used for all the tests.

Understanding these results in a one-by-one basis is rather cumbersome as it implies cross-examining and comparing the results presented separately and would require larger amount of space and effort. That is why we decided to adopt a more integrative representation.

That is, for a given set of algorithms $A_1,\ldots, A_K$, a set of P test problem instances $\Phi_{1,m},\ldots,\Phi_{P,m}$, configured with m objectives, the function $\delta(\cdot)$ is defined as

$$\delta\left(A_i, A_j, \Phi_{p,m}\right) = \begin{cases} 1 & \text{if } A_i \gg A_j \text{ solving } \Phi_{p,m} \\ 0 & \text{in other case} \end{cases} , \tag{14}$$

where the relation $A_i \gg A_j$ defines if A_i is significantly better than A_j when solving the problem instance $\Phi_{p,m}$, as computed by the statistical tests previously described.

Relying on $\delta(\cdot)$, the performance index $P_{p,m}(A_i)$ of a given algorithm A_i when solving $\Phi_{p,m}$ is then computed as

$$P_{p,m}(A_i) = \sum_{j=1;\, j\neq i}^{K} \delta\left(A_i, A_j, \Phi_{p,m}\right). \tag{15}$$

This index intends to summarize the performance of each algorithm with regard to its peers.

Figure 2 exhibits the results computing the performance indexes. Figure 2a represents the mean performance indexes yielded by each algorithm when solving each problem in all of its configured objective dimensions,

$$\overline{P}_p(A_i) = \frac{1}{|\mathcal{M}|} \sum_{m\in\mathcal{M}} P_{p,m}(A_i) \; ; \; \mathcal{M} = \{3, 5, 7, 9\} \; . \tag{16}$$

It is worth noticing that GNG and, particularly, MB-GNG have better overall results than the other algorithms. It is somewhat unexpected that the randomized leader

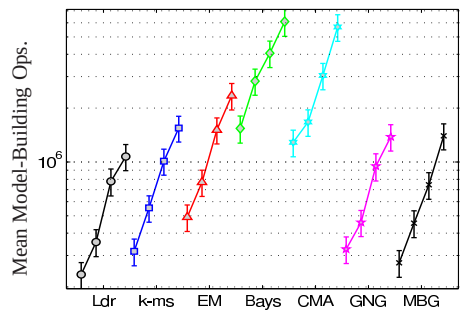


Figure 3: Progression of the mean floating-point CPU operations used by the model-building algorithms as the objective space dimension increases. For each algorithm four points are plotted corresponding to $M = \{3, 5, 7, 9\}$.

and the k-means algorithms do not have a very good overall performance for some problems, like WFG5 and WFG7 for the randomized leader and WFG8 and WFG9 for k-means. A possible hypothesis is that these results may be biased by the three-objective problems, where there are sizable differences compared with the results of the other dimensions.

This situation is clarified in Figure 2b, which presents the mean values of the index computed for each dimension,

$$\overline{P}_m(A_i) = \frac{1}{P} \sum_{p=1}^{P} P_{p,m}(A_i) \ . \tag{17}$$

In this representation, it becomes noticeable that MB-GNG outperforms the other approaches in dimensions larger than three, with the exception of the nine objectives case, in which the original GNG outperforms MB-GNG.

Another key issue when dealing with high-dimensional problems is the computational cost of the algorithms. One simple way of inspecting this point is to compute the number of CPU operations dedicated to model building in each case. Figure 3 summarizes these results. NSGA-II and SPEA2 are not included in the analysis, since they do not perform any model building. The main conclusion in this case is that MB-GNG requires more or less the same amount of resources to yield better results than the other approaches.

6. Final Remarks

In this paper we have discussed an important issue in current evolutionary multi-objective optimization: how to build algorithms that have better scalability with regard to the number of objectives. In particular, we have focused on one promising set of approaches: estimation of distribution algorithms.

We have argued that most of the current approaches do not take into account the particularities of the model-building problem that they are addressing and that, for this reason, they fail to yield results of substantial quality.

Regarding this, we have presented a new model-building algorithm, the MB-GNG, which was assembled in order to meet the requirements of the particular problem.

We have also carried out a set of experiments that showed the points being discussed. The experiments illustrated empirically that MB-GNG outperforms current approaches and is a valid starting point for future advancements in this area. The main aim of this paper is to trigger further studies on this topic and, ultimately, to produce more model-building algorithms.

Acknowledgements

The authors wish to thank the referee and the associate editor assigned to this paper for their comments and suggestions. They helped to substantially improve the paper. They also wish to thank Prof. Elisenda Molina for her assistance in the preparation of the manuscript.

L. Martí, J. García, A. Berlanga and J. M. Molina were supported by projects CICYT TIN2008-06742-C02-02/TSI, CICYT TEC2008-06732-C02-02/TEC, SINPROB, CAM MADRINET S-0505/ TIC/0255 and DPS2008-07029-C02-02. C. A. Coello Coello acknowledges support from CONACyT project 103570.

References

[1] K. Miettinen, Nonlinear Multiobjective Optimization, volume 12 of *International Series in Operations Research & Management Science*, Kluwer, Norwell, MA, 1999.

[2] M. Ehrgott, Multicriteria Optimization, volume 491 of *Lecture Notes in Economics and Mathematical Systems*, Springer, 2005.

[3] T. Bäck, Evolutionary Algorithms in Theory and Practice, Oxford University Press, New York — Oxford, 1996.

[4] J. Branke, K. Miettinen, K. Deb, R. Słowiǹski (Eds.), Multiobjective Optimization, volume 5252 of *Lecture Notes in Computer Science*, Springer–Verlag, Berlin/Heidelberg, 2008.

[5] C. A. Coello Coello, G. B. Lamont, D. A. Van Veldhuizen, Evolutionary Algorithms for Solving Multi-Objective Problems, Genetic and Evolutionary Computation, Springer, New York, second edition, 2007.

[6] R. C. Purshouse, P. J. Fleming, On the evolutionary optimization of many conflicting objectives, IEEE Transactions on Evolutionary Computation 11 (2007) 770–784.

[7] T. J. Stewart, O. Bandte, H. Braun, N. Chakraborti, M. Ehrgott, M. Göbelt, Y. Jin, H. Nakayama, S. Poles, D. Di Stefano, Real–world applications of multiobjective optimization, in: J. Branke, M. Kaisa, K. Deb, R. Słowiǹski (Eds.), Multiobjective Optimization, volume 5252 of *Lecture Notes in Computer Science*, Springer–Verlag, Berlin/Heidelberg, 2008, pp. 285–327.

marti–mb-gng–or-letters.tex 1248 2010-11-24 04:24:46Z lm-deep

[8] J. A. Lozano, P. Larrañaga, I. Inza, E. Bengoetxea (Eds.), Towards a New Evolutionary Computation: Advances on Estimation of Distribution Algorithms, Springer–Verlag, 2006.

[9] M. Pelikan, K. Sastry, D. E. Goldberg, Multiobjective estimation of distribution algorithms, in: M. Pelikan, K. Sastry, E. Cantú-Paz (Eds.), Scalable Optimization via Probabilistic Modeling: From Algorithms to Applications, Studies in Computational Intelligence, Springer–Verlag, 2006, pp. 223–248.

[10] H. Trautmann, T. Wagner, B. Naujoks, M. Preuss, J. Mehnen, Statistical methods for convergence detection of multi-objective evolutionary algorithms, Evolutionary Computation 17 (2009) 493–509. PMID: 19916777.

[11] L. Martí, J. García, A. Berlanga, J. M. Molina, An approach to stopping criteria for multi-objective optimization evolutionary algorithms: The MGBM criterion, in: 2009 IEEE Conference on Evolutionary Computation (CEC 2009), IEEE Press, Piscataway, New Jersey, 2009, pp. 1263–1270.

[12] M. Pelikan, D. E. Goldberg, E. Cantú-Paz, BOA: The Bayesian optimization algorithm, in: W. Banzhaf, J. Daida, A. E. Eiben, M. H. Garzon, V. Honavar, M. Jakiela, R. E. Smith (Eds.), Proceedings of the Genetic and Evolutionary Computation Conference GECCO-1999, volume I, Morgan Kaufmann Publishers, San Francisco, CA, Orlando, FL, 1999, pp. 525–532.

[13] G. Cooper, Computational complexity of probabilistic inference using Bayesian belief networks, Artificial Intelligence 42 (1990) 393–405.

[14] P. Dagum, M. Luby, Approximating probabilistic inference in Bayesian belief networks is NP-hard, Artificial Intelligence 60 (1993) 141–153.

[15] N. Khan, D. E. Goldberg, M. Pelikan, Multi–objective Bayesian optimization algorithm, in: W. B. Langdon, E. Cantú-Paz, K. Mathias, R. Roy, D. Davis, R. Poli, K. Balakrishnan, V. Honavar, G. Rudolph, J. Wegener, L. Bull, M. Potter, A. Schultz, J. Miller, E. Burke, N. Jonoska (Eds.), Proceedings of the Genetic and Evolutionary Computation Conference (GECCO'2002), Morgan Kaufmann Publishers, San Francisco, California, 2002, p. 684.

[16] M. Pelikan, K. Sastry, D. E. Goldberg, Multiobjective hBOA, clustering, and scalability, in: GECCO '05: Proceedings of the 2005 conference on Genetic and evolutionary computation, ACM Press, New York, NY, USA, 2005, pp. 663–670.

[17] J. Ocenasek, Parallel Estimation of Distribution Algorithms, Ph.D. thesis, Faculty of Information Technology, Brno University of Technology, Brno, Czech Republic, 2002.

[18] C. W. Ahn, Advances in Evolutionary Algorithms. Theory, Design and Practice, Springer, 2006.

[19] P. A. N. Bosman, D. Thierens, The balance between proximity and diversity in multiobjective evolutionary algorithms, IEEE Transactions on Evolutionary Computation 7 (2003) 174–188.

[20] J. A. Hartigan, Clustering Algorithms, Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, New York, 1975.

[21] R. Xu, D. Wunsch II, Clustering, IEEE Press Series on Computational Intelligence, Wiley — IEEE Press, New York, illustrated edition, 2008.

[22] M. Costa, E. Minisci, E. Pasero, An hybrid neural/genetic approach to continuous multi-objective optimization problems, in: B. Apolloni, M. Marinaro, R. Tagliaferri (Eds.), Italian Workshop on Neural Neural Nets (WIRN), volume 2859 of *Lecture Notes in Computer Science*, Springer, 2003, pp. 61–69.

[23] Q. Zhang, A. Zhou, Y. Jin, RM–MEDA: A regularity model-based multiobjective estimation of distribution algorithm, IEEE Transactions on Evolutionary Computation 12 (2008) 41–63.

[24] C. Igel, N. Hansen, S. Roth, Covariance matrix adaptation for multi-objective optimization, Evolutionary Computation 15 (2007) 1–28.

[25] C. W. Ahn, R. S. Ramakrishna, Multiobjective real-coded Bayesian optimization algorithm revisited: Diversity preservation, in: GECCO'07: Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation, ACM Press, New York, NY, USA, 2007, pp. 593–600.

[26] J. Shapiro, Diversity loss in general estimation of distribution algorithms, in: Parallel Problem Solving from Nature (PPSN IX), pp. 92–101.

[27] B. Yuan, M. Gallagher, On the importance of diversity maintenance in estimation of distribution algorithms, in: GECCO'05: Proceedings of the 2005 Conference on Genetic and Evolutionary Computation, ACM, New York, NY, USA, 2005, pp. 719–726.

[28] B. Fritzke, A growing neural gas network learns topologies, in: G. Tesauro, D. S. Touretzky, T. K. Leen (Eds.), Advances in Neural Information Processing Systems, volume 7, MIT Press, Cambridge, MA, 1995, pp. 625–632.

[29] T. M. Martinetz, S. G. Berkovich, K. J. Shulten, Neural–gas network for vector quantization and its application to time–series prediction, IEEE Transactions on Neural Networks 4 (1993) 558–560.

[30] A. K. Qin, P. N. Suganthan, Robust growing neural gas algorithm with application in cluster analysis, Neural Networks 17 (2004) 1135–1148.

[31] H. Timm, C. Borgelt, C. Doring, R. Kruse, An extension to possibilistic fuzzy cluster analysis, Fuzzy Sets and Systems 147 (2004) 3–16.

[32] M. J. Schervish, Theory of Statistics, Springer Series in Statistics, Springer, Berlin/Heidelberg, second edition, 1997.

marti–mb-gng–or-letters.tex 1248 2010-11-24 04:24:46Z lm-deep

[33] G. E. P. Box, M. E. Muller, A note on the generation of random normal deviates, Annals of Mathematical Statistics 29 (1958) 610–611.

[34] S. Huband, P. Hingston, L. Barone, L. While, A review of multiobjective test problems and a scalable test problem toolkit, IEEE Transactions on Evolutionary Computation 10 (2006) 477–506.

[35] J. Knowles, L. Thiele, E. Zitzler, A Tutorial on the Performance Assessment of Stochastic Multiobjective Optimizers, TIK Report 214, Computer Engineering and Networks Laboratory (TIK), ETH Zurich, 2006.

[36] W. J. Conover, Practical Nonparametric Statistics, John Wiley & Sons, New York, 3rd edition, 1999.

marti–mb-gng–or-letters.tex 1248 2010-11-24 04:24:46Z lm-deep