A Pareto-Compliant Surrogate Approach for Multiobjective Optimization

Ilya LOSHCHILOV
TAO, INRIA Saclay & LRI,
CNRS UMR 8623
U. Paris Sud, F-91405 Orsay

Marc SCHOENAUER
TAO, INRIA Saclay & LRI,
CNRS UMR 8623
U. Paris Sud, F-91405 Orsay
firstname.lastname@inria.fr

Michèle SEBAG
CNRS, LRI UMR 8623 & TAO,
INRIA Saclay
U. Paris Sud, F-91405 Orsay

ABSTRACT

This paper discusses the idea of using a single Pareto-compliant surrogate model for multiobjective optimization. While most surrogate approaches to multi-objective optimization build a surrogate model for each objective, the recently proposed mono surrogate approach [3] aims at building a global surrogate model defined on the decision space and tightly characterizing the current Pareto set and the dominated region, in order to speed up the evolution progress toward the true Pareto set. This surrogate model is specified by combining a One-class Support Vector Machine (SVMs) to characterize the dominated points, and a Regression SVM to clamp the Pareto front on a single value. The aims of this paper are to identify issues of the proposed approach demanding further study and to raise the question of how to efficiently incorporate quality indicators, such as the hypervolume into the surrogate model.

Categories and Subject Descriptors

I.2.8 [**Computing Methodologies**]: Artificial Intelligence, Problem Solving, Control Methods, and Search

General Terms

Algorithms

Keywords

Multiobjective Optimization, Surrogate Models, Support Vector Machine

1. INTRODUCTION

Surrogate methods have received a particular attention in the realm of Evolutionary Algorithms (EAs), all the more so as EAs are known to require a high number of objective function computations (see e.g. [1] for a survey of surrogate evolutionary optimization). Several types of meta-models have been used (quadratic models, neural networks, Regression Support Vector Machines, kriging or Gaussian Processes). Meta-models can aim at either a global approximation of the objective function, or a local one, focusing on the neighborhood of the best current individuals. The meta-model can be used to replace the objective function for a given number of generations; it can be used to generate new individuals (the optima of the meta-model) from scratch; and it can also be used to filter out unpromising offspring.

Unsurprisingly, Evolutionary Multi-Objective (EMO) algorithms facing even more severe computational issues than single-objective optimization, the use of meta-models has been intensively investigated in the EMO literature (see [2] for a comprehensive survey). Most approaches carry over the single-objective surrogate approach, learning one meta-model for each objective and embedding the meta-models within a standard EMO with little modification [6].

Recently proposed approach [3] aims at building a global mono surrogate model in decision space, characterizing whether an individual belongs to i/ the current Pareto set; or ii/ the dominated region; or iii/ the rest of the decision space. This surrogate model, providing an aggregated perspective on all objective functions simultaneously, is used to guide the search in the vicinity of the current Pareto set, and speed up the population move toward the true Pareto set. This Aggregated Surrogate Model (ASM) is constructed by combining ideas from Regression and One-class SVMs.

Section 2 describes the formulation of the ASM model and gives an overview of the EMO algorithm using ASM, referred to as PARETO-SVM. Section 3 discusses the open issues and Section 4 concludes the paper.

2. PARETO SUPPORT VECTOR MACHINE

2.1 Support Vector Machine

Support vector machines were originally developed for pattern classification and later extended to One-class SVM [4] (the case of one-category data set) and regression, called often Support Vector Regression [5] (Fig. 1 (c-f)).

Given training vectors $x_i \in \mathbb{R}^n$, $i = 1 \ldots \ell$, in two classes, and a vector $y_i \in \mathbb{R}^l$ such that $y_i \in \{-1, 1\}$, Classification SVM solves the following primal problem:

$$\underset{\{w,\ \rho,\ \xi\}}{\text{Minimize}}\ \frac{1}{2}||w||^2 + C\sum_{i=1}^{\ell}\xi_i$$

subject to $\quad y_i(< w, \Phi(x_i) > +\rho) \geq 1 - \xi_i \quad , \quad \xi_i \geq 0$

Its dual form can be solved as quadratic programming

GECCO'10, July 7–11, 2010, Portland, Oregon, USA.

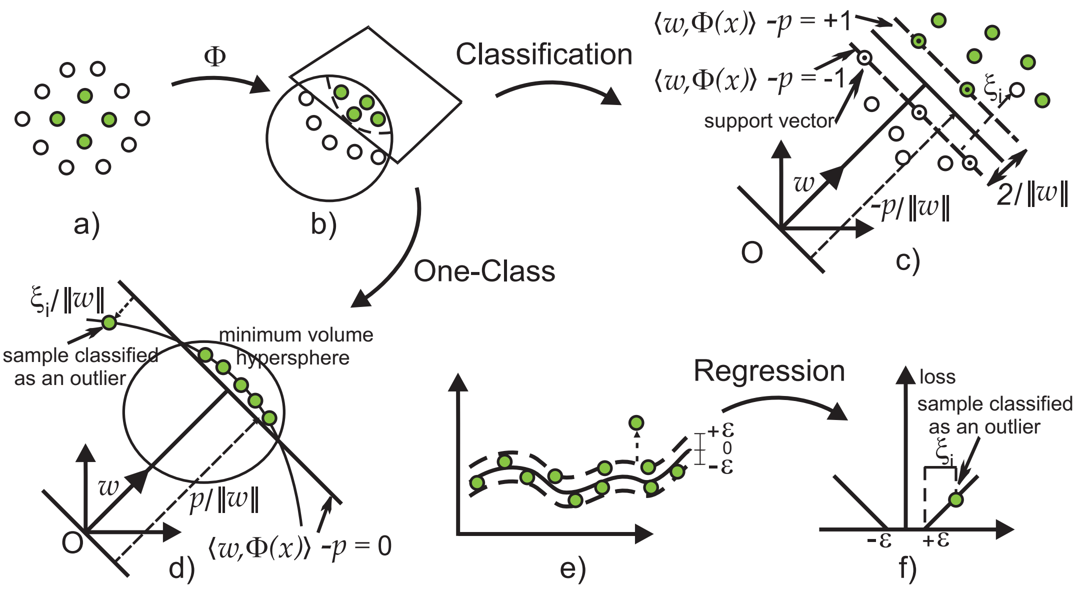
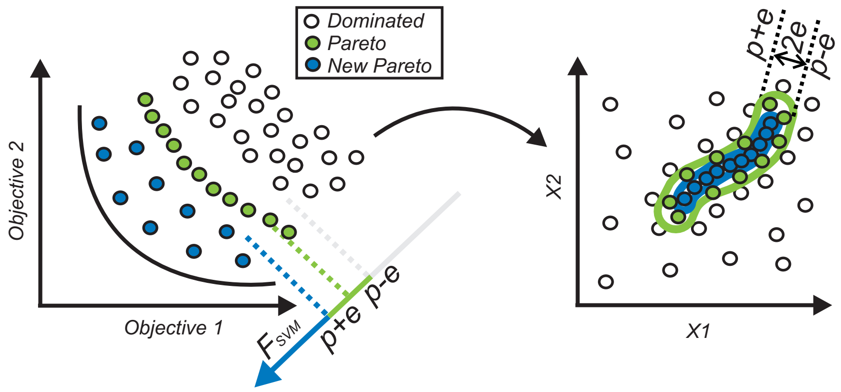


Figure 1: Classification Support Vector Machines (c) is mapping the original linearly non-separable data set (a) into a higher-dimensional feature space (b) by some nonlinear map Φ , where mapped data set can become linearly separable. SVM has been extended to One-class SVM (d) and Support Vector Regression (e,f).



(a) Objective space

(b) Decision space

Figure 2: An idealistic schematic view of the Pareto front, depicting dominated points (white), current Pareto (grey) and new Pareto (black) respectively in objective and decision space.

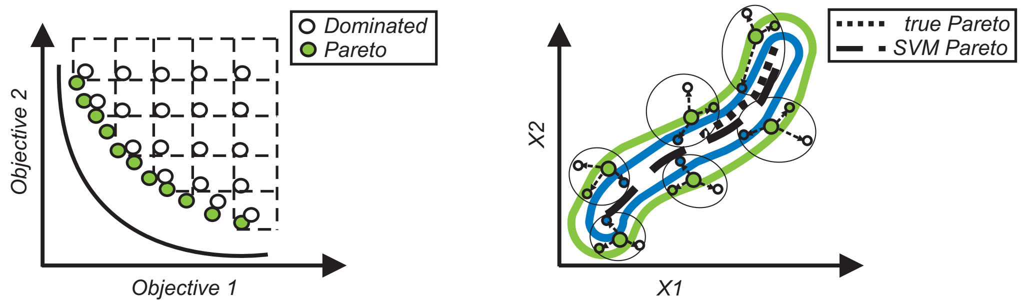


Figure 3: A schematic view of the training data selection in objective space (left) and SVM-informed selection of children from the pool of pre-children in decision space (right).

problem:

$$\underset{\alpha}{\text{Maximize}} \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

subject to $0 \leq \alpha_i < C$, $\sum_{i=1}^{\ell} \alpha_i y_i = 0$
where w is the normal to "separating hyperplane"; $||w||$ is the Euclidean norm of w; ξ are the slack variables introduced for the soft margin case; the constant $C > 0$ determines the trade-off between margin maximization and training error minimization.

The SVM approach, initially aimed at finding linear functions, only computes scalar products of sample points. The so-called kernel trick supports the extension to non-linear functional spaces: the search space X (Fig. 1 (a)) is mapped onto a more expressive space (Fig. 1 (b)) referred to as feature space $\Phi(X)$, where the scalar product $< \Phi(x), \Phi(x') >= K(x, x')$ can be calculated without computing explicitly $\Phi(x)$ or $\Phi(x')$. One example is the Gaussian Radial Basis kernel function (RBF):

$$K(x_i, x_j) = e^{-\|x_i - x_j\|^2 / 2\sigma^2} \tag{1}$$

where σ is a bandwidth parameter.

The decision function $f(x) = sgn(\sum_{i=1}^{\ell} y_i \alpha_i K(x_i, x) + p)$

2.2 Rationale and Assumption

The goal of the present approach is to build a single surrogate model in the decision space, usable to drive the population toward the *true* Pareto set. This surrogate model will be learned from i/ points belonging to the current Pareto set, and ii/ dominated points.

At a given time during the run of an EMOA, the relative position of the Pareto set and the dominated points can be schematically depicted as follows. The situation might be simple in the objective space (Fig. 2.(a)), with the true Pareto front and the dominated region located on the two opposite sides of the current Pareto front. It can be much more intricate in the decision space; Fig. 2.(b) illustrates the case where the true Pareto set (respectively the dominated region) lies within (resp. outside) the interior region of the current Pareto set. Further, the Pareto set can include many disjoint regions in the decision space.

Expectedly, the ASM model discriminates the Pareto set and the dominated region. However, a binary classification approach is ill-suited, in the sense that it would not give any precise indication about where the *true* Pareto set is located. More generally, the Pareto set (true or current) and the dominated points cannot be handled in a symmetrical way: dominated points span over a subspace whereas the Pareto set should better be viewed as a manifold.

It thus comes to map all Pareto points onto a single value ρ (up to some tolerance ϵ); meanwhile, the dominated points would be mapped onto the half space $] - \infty, \rho - \epsilon[$. Such a mapping might actually provide useful indications: expectedly, points mapped onto the half space $[\rho + \epsilon, +\infty[$ would belong to the yet unexplored region, which is bound to contain the *true* Pareto set, and these points could thus be considered promising.

The above constraints on the ASM mapping can be expressed by combining the SVM-regression formulation (mapping each point x onto some target value $f(x)$ up to some tolerance ϵ) and the One-class SVM, mapping a set of points onto a connected interval and thus characterizing the support of the underlying sample distribution. The main difference is that the target value ρ associated to the Pareto points is free in the ASM problem.

2.3 Lagrangian formulation

Let $X \subset \mathbb{R}^d$ denote the search (decision) space and let $x_1 \dots x_m$ denote points in X, with $x_1 \dots x_\ell$ being current Pareto points and $x_{\ell+1}, \dots, x_m$ being dominated points. The sought ASM mapping, noted $\mathcal{F}$ ($\mathcal{F} : X \mapsto \mathbb{R}$), is finally subject to $m + \ell$ constraints: for each $x_i, 1 \leq i \leq \ell$, $\mathcal{F}(x_i)$ must belong to $[\rho - \epsilon, \rho + \epsilon]$ and for each $x_i, \ell < i \leq m$, $\mathcal{F}(x_j)$ must be less than $\rho - \epsilon$.

2.3.1 The primal problem

Using the kernel trick, mapping $\mathcal{F}$ will be defined as a linear function w w.r.t. some feature space $\Phi(X)$:

$$\mathcal{F}(x) = \ < w, \Phi(x) >$$

Then, introducing the usual slack variables $\xi^{(*)}$ (with notations borrowed from [5], $\xi^{(*)}$ represents the $(m + \ell)$-vector made of $(\xi_i^{up})_{i \in [1,\ell]}$, $(\xi_i^{low})_{i \in [1,\ell]}$, and $(\xi_i^{up})_{i \in [\ell+1,m]}$), and given positive constants C and ϵ, the primal problem is:

$$\underset{\{w,\ \xi^{(*)},\ \rho\}}{\text{Minimize}} \frac{1}{2}||w||^2 + C \sum_{i=1}^{\ell} (\xi_i^{up} + \xi_i^{low}) + C \sum_{i=\ell+1}^{m} \xi_i^{up} + \rho \tag{2}$$

subject to

$$< w, \Phi(x_i) > \leq \rho + \epsilon + \xi_i^{up} \quad (i = 1 \dots \ell) \tag{3}$$
$$< w, \Phi(x_i) > \geq \rho - \epsilon - \xi_i^{low} \quad (i = 1 \dots \ell) \tag{4}$$
$$< w, \Phi(x_i) > \leq \rho - \epsilon + \xi_i^{up} \quad (i = \ell + 1 \dots m) \tag{5}$$
$$\xi_i^{up} \geq 0 \quad (i = 1 \dots \ell) \tag{6}$$
$$\xi_i^{low} \geq 0 \quad (i = 1 \dots \ell) \tag{7}$$
$$\xi_i^{up} \geq 0 \quad (i = \ell + 1 \dots m) \tag{8}$$

For the sake of completeness, but due to space limitations, the detailed derivation of the solution is available in annex at `http://sites.google.com/site/paretosvm/`.

2.4 Pareto-SVM Filter Algorithm

This subsection briefly describes the Pareto-SVM Filter Algorithm and some conclusions from the experiments [3].

To efficiently apply the Pareto-SVM learning, the maximum size of training data should be limited, the objective space can be divided regularly into $N_{archive}$ boxes, and only one non-dominated point will be kept in each box (Fig. 3 (left)).

The main difference between general Multiobjective Evolutionary Algorithm (MOEA) and SVM-informed MOEA is the call to the informed operators that replaces the standard call to variation operators, with the ASM as additional argument.

ASM can be used to filter out unpromising offspring. When a variation operator is called, it generates a given number of *pre-children p* for each feature child (Fig. 3 (right)). The value of the surrogate model for all these pre-children is computed, and the operator returns the best one according to those surrogate values. The filtering is more conservative case than direct optimization of ASM model, therefore the potential speed-up is smaller. Thus, the experiments show that two SVM-informed MOEAs with 2-10 pre-children are

1.5-3 times faster than original MOEAs (in terms of number of function evaluations to reach the target hypervolume value on selected bi-objective problems). ASM does not deal with the diversity, therefore for large number of pre-children, the acceleration of optimization can lead to premature convergence.

3. DISCUSSION

3.1 Pareto-SVM formulation

This constrained optimization problem happens to be over-constrained; in such cases, it results in a poor generalization error of the ASM (visible e.g. from its errors on the rest of the Pareto archive). This problem was fixed using an additional k factor, replacing ρ by $k\rho$ in Equation(2). The best k value w.r.t. the ASM generalization error was determined in original algorithm from a preliminary trial, leading to $k = 1$ for one set of problems and $k = -1$ for another set. Probably, a better solution is to reformulate Equation(2):

$$\underset{\{w,\ \xi^{(*)},\ \rho\}}{\text{Minimize}} \frac{1}{2}||w||^2 + C\sum_{i=1}^{\ell}(\xi_i^{up} + \xi_i^{low}) + C\sum_{i=\ell+1}^{m}\xi_i^{up} + \frac{1}{2}D\rho^2$$

Here we want to cluster the Pareto set close to a straight line, and the non-Pareto away from this line, but on one side only. So the problem should stay symmetrical, in particular should be invariant when you change w into -w and ρ into -ρ. But we cannot tell a priori on which side of the line should the non-Pareto be, and choosing arbitrarily one side, as was done in the original formulation, seems to be too much a constraint. In the new formulation, D is an additional problem-dependent parameter, which in some sense makes the learning problem harder.

On-going work aims at understanding this phenomenon +/-ρ, and relating it to the structure of the multi-objective landscape.

3.2 Quality indicator-based Mono surrogate

The original Pareto-SVM approach uses only the information about the dominance relations between points, thus the points from the second and from the 50th non-dominated fronts are the same, they are dominated points. Let's imagine some many-objective problem and the distribution of initial population in the objective space, probably all points of initial population will be non-dominated. Obviously, the original Pareto-SVM approach can not be applied at least for several generations. It seems to be raisonable to use the quality indicator such as the hypervolume to mesure the fitness of solutions, the question is the hypervolume indicator provides enough information for the surrogate model?

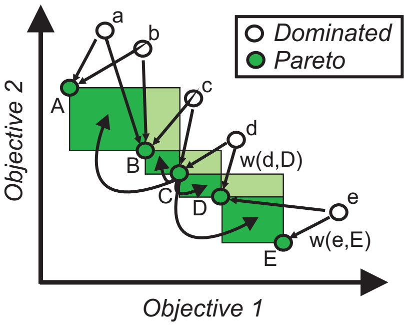


Figure 4: Quality of dominated points as weighted sum of that of nearest Pareto points.

Figure 4 illustrates ten solutions of bi-objective problem. The hypervolume contribution can be used as the fitness function for the Pareto points. But the extreme points will then have infinity contribution, while for other points such as point B the contribution is the volume dominated by B containing no other solution from population. The hypervolume contribution is often used as second-level sorting criterion to eliminate the most crowded points. For each point from non-dominated front the rank value can be defined according to the hypervolume contribution value. But problems arise when we try to estimate the points from the second front. Thus, the point e has the 6th rank, while point E and point C have the first and the 5th rank respectively. But let us suppose that the hypervolume contribution based ranking is very expensive for all non-dominated points. Then, since the hypervolume contribition values of Pareto is known, the fitness of dominated points could be calculated as weighted sum of the fitness function values of the nearest Pareto points in objective space. These weights could also incorporate additional information, such as crowding in decision space.

It is not obvious that the hypervolume contribution should be used as equivalent of fitness function for Pareto points. Probably more complex criterion should be chosen which incorporate the estimation of nearest less crowded region, because even if three Pareto points lie nearly in the same place, but in less crowded region, there is no reason to set a low rank to one of them.

4. CONCLUSION

This paper discussed some open issues related to the idea of using a single Pareto-compliant surrogate model for multi-objective optimization proposed in this conference [3]. This aggregated surrogate model, ASM, enables to guide the offspring generation and speeds up the population move toward the true Pareto set. However, the formulation of the optimization problem still requires a user decision. Furthermore, incorporating other quality indicators when building the surrogate model is appealing, but remains to be done.

5. REFERENCES

[1] Y. Jin. A Comprehensive Survey of Fitness Approximation in Evolutionary Computation. *Soft Computing*, 9(1):3–12, 2005.

[2] J. Knowles and H. Nakayama. Meta-modeling in multiobjective optimization. In J. Branke et al., editor, *Multiobjective Optimization*, number 5252 in LNCS, pages 245–284. Springer Verlag, 2008.

[3] I. Loshchilov, M. Schoenauer, and M. Sebag. A Mono Surrogate for Multiobjective optimization. In J. B. et al., editor, *GECCO'2010*. ACM Press, July 2010. To appear.

[4] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson. Estimating the Support of a High-Dimensional Distribution. *Neural Computation*, 13:1443–1471, 2001.

[5] A. J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, 2004.

[6] I. Voutchkov and A. Keane. Multiobjective Optimization using Surrogates. In I. Parmee, editor, *ACDM'06*, pages 167–175. Institute for People-centred Computation, 2006.