Model-based Methods for Continuous and Discrete Global Optimization

Thomas Bartz-Beielstein, Martin Zaefferer
Model-based Methods for Continuous and Discrete Global Optimization

Thomas Bartz-Beielstein*, Martin Zaefferer

TH Köln, Faculty of Computer Science and Engineering Science, Steinmüllerallee 1, 51643 Gummersbach, Germany

Abstract

The use of surrogate models is a standard method to deal with complex, real-world optimization problems. The first surrogate models were applied to continuous optimization problems. In recent years, surrogate models gained importance for discrete optimization problems. This article, which consists of three parts, takes care of this development. The first part presents a survey of model-based methods, focusing on continuous optimization. It introduces a taxonomy, which is useful as a guideline for selecting adequate model-based optimization tools. The second part provides details for the case of discrete optimization problems. Here, six strategies for dealing with discrete data structures are introduced. A new approach for combining surrogate information via stacking is proposed in the third part. The implementation of this approach will be available in the open source R package SPOT2. The article concludes with a discussion of recent developments and challenges in both application domains.

Keywords: Surrogate, Discrete Optimization, Combinatorial Optimization, Metamodels, Machine learning, Expensive optimization problems, Model management, Evolutionary computation

*This is an extended version of the contribution [1]

*Corresponding author. Phone: +49 2261 8196 6391

Email addresses: thomas.bartz-beielstein@th-koeln.de (Thomas Bartz-Beielstein), martin.zaefferer@th-koeln.de (Martin Zaefferer)
1. Introduction

Model-based optimization (MBO) plays a prominent role in today’s modeling, simulation, and optimization processes. It can be considered as the most efficient technique for expensive and time-demanding real-world optimization problems. Evaluating a cheaper surrogate model instead of the expensive objective function may significantly reduce time, space, and computing costs. Especially in the engineering domain, MBO is an important practice. Recent advances in computer science, statistics, and engineering in combination with progress in high-performance computing provide tools for handling problems, which were considered unsolvable only a few decades ago.

The first part of this article presents a survey of MBO for continuous and discrete global optimization problems. Our goal is to show connections and conceptual differences between these two domains and to discuss properties of state-of-the-art MBO algorithms. Despite its growing relevance, contributions to the discrete domain have been largely disregarded and were listed as open challenges. For example, Simpson et al. [2] present an interesting history of development in the field. Another survey of metamodeling techniques, which focuses on the practitioner’s perspective, is given by Wang and Shan [3]. Both of these surveys mention problems from the discrete domain, but do not discuss them in depth.

The term global optimization (GO) will be used for algorithms and problems where the goal is to find and explore global optimal solutions with complex, multimodal objective functions [4]. Furthermore, we will focus on GO problems which belong to the class of difficult (expensive) black-box functions, i.e., functions for which the analytic form is unknown. Thus, nearly no structural information (e.g., number of local extrema, derivative information) is available. This setting arises in many real-world systems when the explicit form of the objective function \( f \) is not readily available, e.g., if the user has no access to the source code of a simulator. Surrogates are a popular choice, because the time required for building the surrogate is negligible compared to the evaluation of
the real-world function $f$.

The remainder of this article is structured as follows. Section 2 introduces a taxonomy of search algorithms and presents basic definitions. After introducing instance-based stochastic search algorithms, Section 3 describes modeling approaches for stochastic algorithms. We differentiate between models that use a distribution and models that use an explicit surrogate model. Section 4 introduces model-based optimization, which is the first choice for many optimization problems in industry. Problems and algorithms from the discrete, combinatorial domain are then introduced in Section 5. Using two MBO algorithms, namely EvoLS and SPO, recent trends and new developments in MBO are described in Section 6. Finally, a summary and an outlook are given in Section 7, including important challenges in continuous and discrete MBO.

2. Taxonomy and Definitions

We consider the continuous optimization problem given by

$$\text{Minimize: } f(\bar{x}) \quad \text{subject to } \bar{a} \leq \bar{x} \leq \bar{b},$$

where $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is referred to as the objective function and $\bar{a}$ and $\bar{b}$ denote the lower and upper bounds of the search space, respectively. The objective function $f$ is assumed to be an expensive-to-evaluate black-box.

In the discrete case, $\bar{x} \in \mathbb{R}^n$ is not true anymore. Rather, a candidate solution is some discrete data structure, or object. Typical discrete data structures include, e.g., ordinal integers, categorical variables, binary variables, permutations, strings, trees or graphs in general. In most of these discrete cases, bounds $\bar{a}$ and $\bar{b}$ are also not required anymore.

From the algorithmic perspective, this survey focuses on search heuristics, which are mostly implemented using stochastic or random procedures. Deterministic, exact GO algorithms, i.e., algorithms providing theoretical guarantees that the attained solution is the global one within some pre-specified tolerance, are not further discussed [5]. The terms “random” and “stochastic” will be used synonymously in the remainder of this article.
Stochastic search algorithms can further be categorized as instance-based or model-based algorithms [6]. Furthermore, there are basically two model-based approaches: (a) distribution-based models and (b) surrogate models. We consider three important representatives of surrogate model based optimization: (i) Single surrogate based optimization uses one model for accelerating the search process, (ii) multi-fidelity metamodeling uses several models of the same real system and plays an important role in computational fluid dynamics (CFD) and finite element modeling (FEM) based simulation and optimization, and (iii) ensemble surrogate based optimization combines two or more different surrogate models. This categorization (or taxonomy) of GO algorithms is summarized in Figure 1.

Typical situations, which may occur in MBO, are illustrated in Figure 2.
Kriging is a frequently employed type of surrogate model [7]. Figure 2 shows (in situation C) that the Kriging model can be constructed as a surrogate model of a complex real-world process. Furthermore, a Kriging model can be used as a surrogate model of the relatively complex CFD simulation model (as illustrated in situation E). Some authors define a model as a direct abstraction of real-world processes and a surrogate as a second abstraction, which highlights the properties of the model itself. Following these definitions, the Kriging model in situation C is a model, whereas the Kriging model in situation E is a surrogate. As this has the potential of confusion, we will not use this categorization.

Instead, we will use the term “model” for a simplified abstraction of a complex object, whereas the term “surrogate” will be used for the subset of models, which can be represented by an explicit functional relationship [8, 9]. Usually, a data-driven process, which comprehends the following two steps, is used for constructing a surrogate: (i) generation of data through sampling and (ii) mathematical function fitting. The terms “surrogate”, “surrogate model”, and
“metamodel” will be used synonymously in the following.

3. Stochastic Search Algorithms

An iterative search algorithm that uses a stochastic procedure to generate the next iterate is referred to as a stochastic search algorithm. The next iterate can be a candidate solution to the GO or a probabilistic model, where solutions can be drawn from. Stochastic search algorithms are considered robust and easy to implement, because they do not depend on any structural information of the objective function, such as gradient information or convexity. This feature is one of the main reasons for the popularity of stochastic search in the domain of GO.

Instance-based algorithms ([II.1]) maintain a single solution, \( \bar{x} \), or population, \( P(t) \), of candidate solutions. The iteration or time step is denoted as \( t \). The construction of new candidate solutions depends explicitly on the previously generated solutions. Simulated annealing [10], evolutionary algorithms (EAs) [11, 12], and tabu search [13] are prominent representatives of this category. The key elements of instance-based algorithms are shown in Algorithm 1. This pseudo code focuses on the fundamental structure of the algorithm and skips some implementation details, e.g., \( P(t) \) represents the population at time step \( t \) and the corresponding function values.

Algorithm 1 Instance-based algorithm

1: \( t = 0 \). \( P(0) = \text{SetInitialPopulation}() \).
2: Evaluate\((P(t))\) on \( f \).
3: while not TerminationCriterion() do
4: Generate new solutions \( P'(t) \) according to random mechanism.
5: Evaluate\((P'(t))\) on \( f \).
6: Select the next population \( P(t + 1) \) from \( P(t) \cup P'(t) \).
7: \( t = t + 1 \).
8: end while
4. MBO: Model-based Algorithms

Model-based optimization algorithms ([II.2]) generate a population of new candidate solutions \( P'(t) \) by exploiting a model (surrogate or distribution). The model reflects structural properties of the underlying true function \( f \). Model-based optimization algorithms are based on the idea that by adapting the model, the search is directed into regions with improved solutions.

One of the key ideas in MBO is the replacement of expensive, high fidelity, fine grained function evaluations, \( f(\bar{x}) \), with evaluations, \( \hat{f}(\bar{x}) \), of an adequate cheap, low fidelity, coarse grained model, \( M \). An additional benefit can be seen in the smoothening property of the cheap model, because it uses an approximation of the possibly noisy data [2].

This section describes two different MBO approaches: (i) distribution based ([II.2.1]) and (ii) surrogate-model based optimization ([II.2.2]).

4.1. Distribution-based Approaches

If the metamodel is a distribution, the most basic form of an MBO can be implemented as shown in Algorithm 2:

\begin{algorithm}
\caption{Distribution-based algorithm}
1: \( t = 0 \). Let \( p(t) \) be a probability distribution.
2: while not TerminationCriterion() do
3: Randomly generate a population of candidate solutions \( P(t) \) from \( p(t) \).
4: Evaluate(\( P(t) \)) on \( f \).
5: Generate updated distribution \( p(t+1) \) with population (samples) \( P(t) \).
6: \( t = t + 1 \).
7: end while
\end{algorithm}

Distribution-based algorithms generate a sequence of iterates (probability distributions) \( \{p(t)\} \) with the hope that

\[ p(t) \to p^* \text{ as } t \to \infty, \]

7
where $p^*$ is a limiting distribution, which assigns most of its probability mass to the set of optimal solutions. So it is the probability distribution (as opposed to the candidate solutions in instance-based algorithms) that is propagated from one iteration to the next.

*Estimation of distribution algorithms* (EDA) are well known distribution-based algorithms, which became popular in the field of EAs. Classic EAs vary candidate solutions (also called individuals) with variation operators, such as mutation and recombination. In EDAs, variation operators were replaced by a distribution based procedure: the new population of candidate solutions is generated according to the probability distribution induced or estimated from the promising candidate solution from the current population. Larraaga and Lozano [14] review different ways of using probabilistic models as EDA instantiations.

Although distribution-based approaches play an important role in GO, they will not be discussed in detail in this article. The reader is referred to Hauschild and Pelikan [15], who discuss advantages and outline many of the different types of EDAs. In addition, Hu et al. [16] present recent approaches and a unified view on distribution-based approaches. We will concentrate on surrogate model-based approaches, which have their origin in statistical design and analysis of experiments, especially in response surface methodology [2].

4.2. *Surrogate Model-based Approaches*

In general, surrogates are used, when the determination of a process outcome is very expensive or time consuming. Surrogates imitate the behavior of the real model as closely as possible, while being computationally cheaper to evaluate. The surrogate model is also known as a response surface, metamodel, approximation, coarse grained, or simply the *cheap* model. First approaches were developed for stochastic physical experiments and used traditional design of experiment methods. In the late 1990s, deterministic computer experiments gained popularity. This development shifted the focus of surrogate modeling to the *design and analysis of computer experiments* (DACE) approach [17].
Three variants of surrogate model-based optimization (SBO) became popular in the last years: (i) single surrogate (category [II.2.2.1]), (ii) multi-fidelity approximation ([II.2.2.2]), and (iii) ensemble surrogate ([II.2.2.3]) approaches. The first approach uses one model, the second one uses several instances with different parameterizations of the same model, and the third approach uses several different models.

4.2.1. Single Models

Simple surrogate models are constructed with a data-driven approach. They can be refined by integrating additional points or domain knowledge, e.g., constraints, into the surrogate.

A wide range of surrogates was developed in the last decades. Classical regression models such as polynomial regression or response surface methodology [18], support vector machines (SVM) [19], artificial neural networks [20], radial basis functions [21], or Gaussian process (GP) models, which are sometimes also referred to as DACE or Kriging [22, 23, 24, 25, 26] are the most prominent modeling approaches.

Due to this overwhelming number of approaches, it is desirable to find a good starting point for exploring this important field. Fortunately, Forrester et al. [7] present a comprehensible introduction to SBO with several examples. Furthermore, Wang et al. [3] list popular metamodeling techniques and the related components, such as experimental design, sampling methods, metamodels, and model fitting techniques.

Surrogate-assisted evolutionary algorithms are a popular variant of SBO. These are evolutionary algorithms that decouple the evolutionary search and the direct evaluation of the objective function. A good introduction and a comprehensive survey is presented by Jin [27, 28].

Model-based optimization can be improved by integrating a sequential strategy, which enables a model refinement during the optimization process. Additional points, which are evaluated on the expensive function, $f$, can be used for building the surrogate $M$. As a consequence of this refinement, the surrogate
$M$ depends on the time step, i.e., $M(t)$. A minimalistic surrogate model-based optimization (SBO) algorithm is shown in Algorithm 3.

**Algorithm 3** Surrogate model based optimization (SBO) algorithm

1: $t = 0$. $P(t) = \text{SetInitialPopulation}()$.  
2: Evaluate($P(t)$) on $f$.  
3: while not TerminationCriterion() do  
4:   Use $P(t)$ to build a cheap model $M(t)$.  
5:   $P'(t + 1) = \text{GlobalSearch}(M(t))$.  
6:   Evaluate($P'(t + 1)$) on $f$.  
7:   $P(t + 1) = P(t) \cup P'(t + 1)$.  
8:   $t = t + 1$.  
9: end while

**Adaptive sampling**, a well-known refinement strategy, proceeds as follows: An initial model, which uses a limited amount of sample points from the expensive objective function, is refined by identifying new points, so-called infill points. In the selection of adequate infill points, two conflicting goals have to be satisfied. The infill points can be selected with respect to

- exploration, i.e., improving the model quality (related to the model $M$) or
- exploitation, i.e., improving the optimization and determining the optimum (related to the objective function $f$).

Adaptive sampling tries to find a balance between exploration and exploitation. A popular adaptive sampling method is expected improvement (EI) [29, 30], which is discussed in Jones’ seminal paper [31].

MBO methods have been found to be useful in many applications. For example, simulation-based design of complex engineering problems, e.g., structural design of vehicles, ships, or airplanes use CFD/FEM methods. The solvers require a large number of computer simulations to guarantee an exact solution. Hence, simulation-based design is one of the most popular and successful application areas for MBO.
There are several publications that describe simulation-based design MBO applications in aerospace design [32]. A combination of a genetic algorithm and neural networks for aerodynamic design optimization is suggested by Hajela et al. [33]. The development of effective numerical methods for managing the use of approximation concepts in optimization for a 31-variable helicopter rotor design, which was part of a collaboration between Boeing, IBM, and Rice University, is described by Booker et al. [34, 35]. Giannakoglou [36] discusses an aerodynamic shape design problem. A surrogate-assisted evolutionary optimization framework, which is applied to an airfoil shape optimization problem using computational fluid dynamic is presented by Zhou et al. [37].

Obviously, the huge variety of models results in many innovative approaches for combining successful methods and improving existing optimization algorithms. Ratle [38] creates an approximate model of the fitness landscape using Kriging interpolation to accelerate the convergence of EAs. Jin and et al. [39] investigate the convergence property of an evolution strategy (ES) with neural network based fitness evaluations. Emmerich et al. [40] introduce a surrogate-assisted ES. They present results from classical test functions (artificial landscapes) and from an airfoil-shape optimization problem. A significant reduction in the number of function evaluations was observed in both domains. Jin and Sendhoff [41] use clustering techniques and neural networks ensembles to reduce the number of function evaluations. Branke and Schmidt [42] propose not to evaluate every candidate solution (individual), but to just estimate the objective function value of some of the individuals. The reduction in the number of function evaluations is obtained by estimating an individual’s function value on the basis of previously observed objective function values of neighboring individuals. Zhou et al. [37] present a surrogate-assisted EA framework, which uses computationally cheap hierarchical surrogate models constructed through online learning to replace the exact computationally expensive objective functions during evolutionary search.

Although the application of surrogate modeling techniques has progressed remarkably in the past last decades, the question remains “How far have we
really come?” This issue is addressed by Simpson et al. [2]. They conclude that while methods in the field have improved, the issues and challenges mainly remain the same: dimensionality, computational complexity, noisy data, handling of mixed continuous / discrete data and model validation (cf. Section 7). They attribute this to the fact that the complexity and difficulty of problems increased simultaneously. Better tools (algorithms as well as hardware) allow to handle more difficult problems.

4.2.2. Multi-fidelity Models

In addition to single surrogates, multi-fidelity modeling methods were developed. Multi-fidelity modeling (or synonymously multi-fidelity metamodeling) uses several models of the same real system, where each model has its own degree of detail representing the real process. Sun et al. [43] describe a multi-fidelity optimization approach for sheet metal forming process. Further examples of multi-fidelity modeling are presented by Ulaganathan et al. [44]. The authors analyze the performance of Kriging [31] when multi-fidelity gradient data is introduced along with multi-fidelity function data to approximate black-box simulations.

Koziel et al. [45] present a methodology for fast multi-objective antenna optimization with co-Kriging. Co-Kriging is an extension of Kriging, which uses the correlations between the models of various fidelities, so that cheap- and expensive simulation data can be combined into one surrogate model [46, 47]. Co-Kriging-based sequential design strategies are presented by Le Gratiet and Cannamela [48]. The authors simulate a spherical tank under internal pressure. Further applications from the water industry are published by Razavi et al [49].

Tuo et al. [50] proposed a finite-element analysis with its mesh density as the tuning parameter. A problem in casting simulation is used to illustrate this approach. Kleijnen [51] presents an overview of the most recent approaches in simulation practice, including multi-fidelity modeling.
4.2.3. Ensembles

The model building and selection process is crucial for the effectivity and efficiency of SBO. The EI approach, which was introduced in Section 4.2.1, handles the initialization and refinement of a surrogate model, but not the selection of the model type. The popular efficient global optimization (EGO) algorithm uses a Kriging model, because Kriging inherently determines the prediction variance, which is necessary for the EI criterion [30]. But there is no proof that Kriging is the best choice. Alternative surrogate models, e.g., regression trees, support vector machines, or lasso and ridge regression may be better suited, depending on the problem. For example, Müller and Shoemaker [52] reported that Kriging models and combinations including Kriging performed in general worse than when RBF models are used.

An a priori selection of the best suited surrogate model is often impossible in the framework treated in this article, because of the black-box setting described in Section 1.

Regarding the model choice, the user can decide whether to use (i) a single, global model during the optimization, i.e. category [II.2.2.1] in Figure 1, (ii) multi-fidelity models, i.e., category [II.2.2.2], or (iii) multiple models, \( M_i, i = 1, 2, \ldots, p \), which is category [II.2.2.3]. Ensembles comprehend an aggregate of several models. In contrast to multi-fidelity models, ensembles do not necessarily combine models of different fidelity. Classical ensemble methods rather combine data driven models of similar or identical fidelity. One example are random forests, which represent an ensemble of simple tree-based models.

To reduce the computing time, an ensemble of different, possibly local, models can be generated and evaluated in parallel. Each model may employ the same candidate solutions (from the population \( P \)) and the corresponding observations from expensive function evaluations. Haftka et al. [53] present a review of surrogate methods that provide parallelization.

Since each of the \( p \) surrogates proposes one or even more different candidate solutions, a selection and/or a combination mechanism is required. Convention-
ally, surrogate models are assessed and chosen according to their estimated true error [54, 55, 2]. Generally, attaining a surrogate model that has minimal error is the desired feature. The mean absolute error (L1 norm), the mean square error (MSE) (or its pendant the root mean square error (RMSE)) are commonly used as performance metrics. Methods from statistics, statistical learning [56], and machine learning [57], such as the simple holdout approach, cross-validation, and the bootstrap are also important in this context.

Several selection and combination mechanisms for surrogates were developed in the last years. A simple approach determines the best model, i.e., the model with the smallest prediction error, and determines the next candidate solution based on that model. Alternatively, candidate solutions from several models can be combined. Zerpa et al. [58] use multiple surrogate models and build an adaptive weighted average model of the individual surrogates. Goel at al. [59] explore the possibility of using the best surrogate model or a weighted average surrogate model instead of one single model. Model quality, i.e., the errors in surrogates, is used to determine the weights assigned to each model. Sanchez et al. [60] present a weighted-sum approach for the selection of model ensembles. The models for the ensemble are chosen based on their performance and the weights are adaptive and inversely proportional to the local modeling errors. Tenne and Armfield [61] propose a surrogate-assisted memetic algorithm which generates accurate surrogates using multiple cross-validation tests.

Huang et al. [62] use several simulation models for a semiconductor manufacturing system. They propose an ordinal transformation to utilize the results from several cheap models. The function values of all solution candidates is evaluated on every cheap model and the individuals are ranked. The authors observe that despite the big bias in the results from the cheap models, the relative order among solutions is actually quite accurate. This order can be used to accelerate the selection process in EAs significantly. To reduce variability and bias in the results from the cheap models, the authors apply an optimal computing budget allocation scheme.

Multiple models can also be used to partition the search space. The treed
Gaussian process approach uses regression trees to partition the search space into separate regions and to fit local GP surrogates in each region [63]. Nelson et al. [64] propose an algorithm, which creates a tree-based partitioning of an aerodynamic design space and employs independent Kriging surfaces in each partition. Couckuyt et al. [65] propose to combine an evolutionary model selection (EMS) algorithm with the EI criterion in order to dynamically select the best performing surrogate model type at each iteration of the EI algorithm.

5. Surrogate Models in Discrete Optimization

Compared to their frequent use for real-valued problem domains, surrogate model driven approaches are relatively scarce in combinatorial or mixed optimization [28]. Discrete problems deal, e.g., with ordinal integers, categorical (qualitative) variables, permutations, strings, trees, or other graph based data structures. They may be mixed among themselves, or mixed with continuous variables. Ordinal integer variables can often be handled quite similarly to real valued variables. Others, like trees, are too complex to be easily represented by numeric vectors.

Few expensive, real-world combinatorial optimization problems have been published, e.g., in the engineering domain [66, 67, 68, 69], bioinformatics [70], or data science [71]. Not all of them make use of surrogate models. This scarcity is unlikely due to a lack of problems in this field. Rather, the availability of suitable methods (i.e., surrogate models, category [II.2.2]) is not well known or these methods are not easily accessible to experts in potential application domains. Thus, we provide a survey of surrogate modeling methods for combinatorial, discrete problems. The general taxonomy for these methods is the same as for the continuous case, which was illustrated in Figure 1.

5.1. Strategies for Dealing with Discrete Structures

Table 1 presents a tabulated overview of the literature on discrete, surrogate model-based optimization. This overview presents important stepping stones
Table 1: Overview of surrogate models in combinatorial optimization. Column *data* lists data type: Mixed (mix), ordinal (ord), categorical (cat), binary (bin), permutation (per), signed permutation (-per), trees (tre), other (oth). Column *s* lists the strategy, see Section 5.1. Column *dim* lists the dimensionality of the problem, where applicable. Abbreviations introduced in the table: *genetic algorithm* (GA), *non-dominated sorting genetic algorithm II* (NSGA2), *simulated annealing* (SA), *artificial neural networks* (ANN), *ant colony optimization* (ACO), *multi-start local search* (MLS). This table is continued on page 17.

<table>
<thead>
<tr>
<th>data</th>
<th>STR</th>
<th>model</th>
<th>optimizer</th>
<th>cost</th>
<th>budget</th>
<th>dim</th>
<th>topics</th>
<th>ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>mix, cat, ord</td>
<td>1,3</td>
<td>Kriging, Tree</td>
<td>visual,</td>
<td>high</td>
<td>≤ 100</td>
<td>2, 9</td>
<td>parameter tuning</td>
<td>[72]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>statistical</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mix, ord, cat</td>
<td>6</td>
<td>RBFN</td>
<td>ES</td>
<td>low /</td>
<td>560 /</td>
<td>15</td>
<td>benchmark, medical image</td>
<td>[73]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>~high</td>
<td>280</td>
<td>23</td>
<td>analysis</td>
<td></td>
</tr>
<tr>
<td>mix, ord, cat</td>
<td>3,6</td>
<td>Random Forest, Kriging</td>
<td>MLS</td>
<td>~high</td>
<td>-</td>
<td>4.76</td>
<td>algorithm tuning</td>
<td>[74]</td>
</tr>
<tr>
<td>mix, bin, cat</td>
<td>6</td>
<td>RBFN + cluster</td>
<td>GA</td>
<td>low</td>
<td>2,000</td>
<td>12</td>
<td>benchmark, chemical industry</td>
<td>[75]</td>
</tr>
<tr>
<td>mix, ord, cat</td>
<td>4</td>
<td>SVM</td>
<td>NSGA2</td>
<td>?</td>
<td>2,000</td>
<td>10</td>
<td>FEM, multi criteria</td>
<td>[76]</td>
</tr>
</tbody>
</table>

and interesting applications in the field, thus showcasing the development. For a more extensive table we refer to the tabular overview in the supplemental material of this article\(^1\). In Table 1, the employed modeling strategies, model types, optimizers, and problem-related details are specified. The table lists works on (mixed) integer problems, binary representations, permutation problems, tree structures, and other representations.

To deal with modeling in combinatorial search spaces, a set of six *strategies* (STR) can be identified in the literature, which are referenced in column “STR” of Table 1:

\(^1\)The tabular overview will be kept up-to-date on the second author’s home page [https://martinzaefferer.de/?page_id=134](https://martinzaefferer.de/?page_id=134)
Table 1: continued

<table>
<thead>
<tr>
<th>data</th>
<th>STR</th>
<th>model</th>
<th>optimizer</th>
<th>cost</th>
<th>budget</th>
<th>dim.</th>
<th>topics</th>
<th>ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>bin</td>
<td>1/3</td>
<td>ANN</td>
<td>SA</td>
<td>high</td>
<td>?</td>
<td>16</td>
<td>real world, pump positioning</td>
<td>[77]</td>
</tr>
<tr>
<td>bin</td>
<td>6</td>
<td>RBFN</td>
<td>GA</td>
<td>low</td>
<td>dim^2</td>
<td>10-25</td>
<td>NK-Landscape</td>
<td>[78]</td>
</tr>
<tr>
<td>-per</td>
<td>2</td>
<td>custom</td>
<td>brute force</td>
<td>high</td>
<td>28</td>
<td>6</td>
<td>real world: weld sequence</td>
<td>[66]</td>
</tr>
<tr>
<td>per</td>
<td>6</td>
<td>RBFN</td>
<td>GA</td>
<td>low</td>
<td>100</td>
<td>30-32</td>
<td>benchmark</td>
<td>[79]</td>
</tr>
<tr>
<td>per</td>
<td>6</td>
<td>Kriging</td>
<td>GA</td>
<td>low</td>
<td>100</td>
<td>12-32</td>
<td>benchmark</td>
<td>[80]</td>
</tr>
<tr>
<td>per</td>
<td>6</td>
<td>Kriging</td>
<td>ACO</td>
<td>low</td>
<td>100 - 1,000</td>
<td>50-100</td>
<td>benchmark, tuning</td>
<td>[81]</td>
</tr>
<tr>
<td>tre</td>
<td>6</td>
<td>RBFN</td>
<td>GA</td>
<td>low</td>
<td>100</td>
<td></td>
<td>symbolic regression</td>
<td>[82]</td>
</tr>
<tr>
<td>tre</td>
<td>5,6</td>
<td>k-NN</td>
<td>GA</td>
<td>high</td>
<td>30,000</td>
<td></td>
<td>phenotypic similarity, genetic programming</td>
<td>[83]</td>
</tr>
<tr>
<td>tre</td>
<td>5</td>
<td>Random Forest</td>
<td>GA</td>
<td>low</td>
<td>15,000</td>
<td></td>
<td>benchmark, genetic programming</td>
<td>[84]</td>
</tr>
</tbody>
</table>

STR-1 The naive approach: As long as the data can still be represented as a vector (binary variables, integers, categorical data, permutations) the modeling technique may simply ignore the discrete structure, and work as usual. A potential drawback of this approach is, that the model’s input space may have large areas of redundancy. Or else, this approach may create large areas of infeasible solutions. Depending on the optimizer, this may deteriorate performance compared to more sophisticated approaches.

STR-2 Custom modeling: A specific modeling solution is tailored to fit the needs of a certain application. On the one hand, this procedure can be very efficient, because it integrates significant prior knowledge into the model. On the other hand, it may introduce bias and may be hard to transfer to other applications or data structures. This approach is not applicable for true black-box problems.
STR-3 Inherently discrete models: Some models already are discrete in their own design. One example are tree-based models, like regression trees or random forests. On the one hand, these models are easy to use, because no or only minor adaptations are necessary. On the other hand, this strategy may fail if the discrete structures become more complex (e.g., trees or other graph structures). Also, such models may not always provide desired features, e.g., the nice properties derived from the uncertainty estimates of a Kriging model.

STR-4 Mapping: Often, discrete variables or structures may be mapped to a more easily handleable representation. Examples for this approach are the random key mapping for permutations or dummy variables for categorical variables. Similarly to strategy STR-1, this approach may introduce redundancy or infeasibility into the data structure. Establishing reasonable mappings becomes harder for data structures with increasing complexities.

STR-5 Feature extraction: Instead of directly modeling the relation between an object (or its representation) and its quality, it is possible to calculate real-valued features of the objects. For example, some features of a tree or graph can be extracted (path lengths, tree depths, etc.). These numeric features can then be modeled with standard techniques.

STR-6 Similarity-based modeling: Where available, measures of (dis)similarity may be used to replace continuous measures that are, e.g., employed in similarity-based models like $k$-nearest neighbor ($k$-NN), support vector machines (SVM), radial basis function networks (RBFN), or Kriging. While this approach is potentially very powerful, a drawback is the requirement of proper measures. This may be problematic if these measures have to fulfill further requirements, like definiteness.

The presented six strategies are not necessarily mutually exclusive. Depending on the point of view, a mapping approach can be interpreted as similarity-
based approach, or vice versa. Or else, features may be extracted and used for modeling, while at the same time applying some inherently discrete approach to the data. Thus, some methods may either combine several strategies, or else, can be classified as belonging to several strategies.

None of the six strategies can be broadly preferred to the others. Even the naive approach may be adequate, if the problem is sufficiently simple. Problem type and application restrictions will govern the selection of a suitable approach. The subsequent sections present key characteristics of strategies considered by the authors to be most relevant.

5.2. Custom Models

One way of using surrogate-models for combinatorial problems is to employ customized, application specific solutions (STR-2). An example is the work by Voutchkov et al. [66]. They optimize a weld sequence, which is represented as a signed permutation, and find a near-optimal welding sequence by evaluating only 28 out of 46,080 possible sequences. The surrogate model replaces an expensive FEM by estimating the influence of each individual element in the weld sequence, based on the observations made in previously tested sequences. In addition to the function values, the surrogate also exploits intermediate results that reflect the impact of individual sequence-elements, depending on their position in the weld sequence. Exploiting these intermediate results is a clear advantage over more simple, function-value driven approaches. On the other hand, the applicability of this model is restricted to this specific setup and cannot be easily transferred to completely different application areas. The surrogate modeling approach in [66] has since been adapted and applied to a girth weld path optimization problem by Asadi and Goldak [67]. Due to the nature of their problem (number of weld sub-passes, rotational symmetry) only 48 sequences are possible. They evaluated 15 of these 48 sequences to find an optimal solution. Nguyen et al. [85] propose to use simplifications of the target function (a simulation model for job shop scheduling) as a surrogate model in genetic programming. These applications show that custom models are beneficial, if
domain knowledge is available.

5.3. Mapping

As stated in Section 5.1, a frequently chosen approach is mapping from the more complex, discrete space to another more easy to handle space (STR-4). Classical regression provides an establishes mapping approach for dealing with discrete, categorical parameters, which is based on dummy variables and contrasts. A categorical variable can be mapped to a set of dummy variables, where each dummy represents a single level of the original variable. This is the same as the one-hot encoding, i.e., only one of several bits can be 1. The $n$ levels of the categorical variable are mapped to $n$ binary variables. The binary variables are still discrete, but can be handled by standard regression approaches. Else, one can map to $(n - 1)$ binaries: The one missing level is represented by all dummy variables being zero. This is essentially a contrast. Contrast variables represent a comparison between different levels. Clearly, both approaches increase the number of model parameters. Dummy variables or contrasts are probably among the most frequently applied methods to deal with categorical variables, e.g., in the context of the response surface methodology [86].

Another example of a mapping approach is the random key mapping developed by Bean [87]. Random keys are vectors of real numbers from a fixed interval, e.g., the interval $[0, 1]$. To map from this real parameter space to permutation representations, the random key values are simply sorted. Random keys were originally developed to enable the application of continuous, model-free genetic algorithms to permutation problems. They have since been used in probabilistic models which are part of EDAs, see also Section 4.1 and 5.6. Notable drawbacks are redundancy in the real valued parameter space as well as the non-bijection character of the mapping.

5.4. Feature Extraction

Of the few approaches that deal with modeling for tree-representations (symbolic regression, genetic programming), many use feature extraction for modeling. Hildebrandt and Branke [83] extract features of the phenotypic behavior
of evolved dispatching rules for job shop scheduling. Here, these features are based on the rule decisions for a small set of reference situations. The Euclidean distance is computed on these features and is then employed in a $k$-NN surrogate model. From a different point of view, this process can as well be seen as a part of the (dis-)similarity calculation, and hence part of strategy STR-6, rather than STR-5. Hildebrandt and Branke compare the phenotypic approach to a genotypic distance measure: the structural Hamming distance. The phenotypic feature extraction approach has since also been investigated by Nguyen et al. [88], who improved handling of replications and the selection scheme.

Pilat and Neruda [84] take a slightly different approach. They extract features from the genotype of candidate solutions, e.g., the depth of the tree, summarizing statistics of numeric constants in the tree, or the cardinality of certain arguments in the tree.

Some of these recent approaches use a rather simple surrogate model, i.e., $k$-NN [83, 88]. This is tightly linked to the comparatively high evaluation budgets (tens of thousands). More complex models might become too expensive when data-sizes grow to these dimensions. On the other hand, using such large budgets may be necessary due to the extremely large search spaces that genetic programming is typically dealing with. Compared to the more common, model-free genetic programming approaches, few tens of thousands of evaluations are in fact a small budget.

5.5. (Dis)similarity based models

A promising approach that recently gained more traction is similarity-based modeling (STR-6). Fonseca et al. [89] defined similarity-based models as models that keep a memory of solutions and estimate the performance of new samples by comparing them to that memory. Fonseca et al. list fitness inheritance [90], fitness imitation [91, 27] and $k$-NN [92] as examples. They test a genetic algorithm supported by a $k$-NN model on a set of numerical, continuous test functions. Bernardino et al. [93] perform similar tests with artificial immune systems. In both cases Hamming and Euclidean distances are used as mea-
sures of dissimilarity, showing that this approach does not depend on a specific measure.

However, the $k$-NN model is not able to predict whether or not a candidate solution will provide improvement over existing solutions. For that purpose, another set of similarity or distance based models is of interest: RBFN, SVM, and Kriging (Gaussian process regression). Hemker [94] describes in his Ph.D. thesis an approach to deal with continuous and ordinal integer variables, applied to electrical engineering and water resource management problems. Their approach is based on Kriging, essentially treating all parameters with the same distance function, hence not directly applicable to categorical parameters or more complex representations. Li et al. [73] proposed an adaptation of RBFNs. These adaptations are based on a weighted distance measure, replacing the usual distance measure employed in RBFN. Their distance-based RBFN model was tested with mixed integer optimization problems. It has since been applied to a multi objective building design optimization problem by Brownlee and Wright [95]. Mixed optimization problems also arise in algorithm tuning. In this context, Hutter [74] also used a Kriging model with a Hamming distance based correlation function to handle categorical variables. A combination of Kriging and tree-based models called treed Gaussian processes has been used by Swiler et al. [96]. They applied a treed Gaussian process model, a Kriging model, and a smoothing spline technique to build surrogate models for mixed integer problems. Coelho et al. [97] and Herrera et al. [76] applied a kernel-based regression method to mixed-variable optimization problems. A different approach with RBFs for the case of a mixed (discrete and continuous) optimization problem is taken by Bajer and Holena [75]. Instead of creating one RBFN based on a weighted distance between candidate solutions they use Hamming distance to cluster the discrete variables, then fit a standard RBFN with the continuous variables for each cluster.

Moraglio and Kattan [78] adapted an RBFN to arbitrary distance measures to model arbitrary combinatorial optimization problems. Their approach has also been applied to quadratic assignment problems [79]. A similar conceptual
extension for Kriging was investigated by Zaefferer et al. [80]. Here, Kriging-based EGO [30] showed positive results when applied to combinatorial problems. Zaefferer et al. [98] also showed that maximum likelihood estimation (MLE) can be used to efficiently choose one measure from a set of distance measures.

An investigation by Cáceres et al. [81] reports negative results for the application of EGO to permutation problems. There, an ant colony optimization algorithm was not outperformed by the Kriging-based variant of the same algorithm. Smith et al. [71] describe a study on extremely high-dimensional test instances, employing RBFN models. While the employed models showed some promise, a proposed ensemble of models performed poorly.

The above modeling approaches use distances and kernels in a standard, straight-forward way. However, it is often important to consider if either the employed distances are conditionally negative semi-definite (CNSD) or positive semi-definite (PSD). Definiteness is a frequent requirement for modeling methods based on distances or kernels. Dealing with the possible indefiniteness of a function is hence of utmost concern. In fact, lack of definiteness is one possible source of some previously observed numerical problems, e.g., in [80, 98]. A recent study by Zaefferer and Bartz-Beielstein [99] dealt with the issue of definiteness in the context of Kriging based optimization. They transfer and extend methods from the field of SVMs, which were previously used for machine learning with structured data, c.f., the survey by Schleif and Tino [100]. While, to the best of our knowledge, SVMs have rarely been used as surrogates in combinatorial optimization, they have been applied to learning problems with combinatorial data (see e.g., [101, 102, 103]). Hence, they are a promising choice.

Most of the above references make use of distance measures in genotype space. For the case of genetic programming, Hildebrandt and Branke [83] show, that a distance in phenotype space may be an excellent choice, see also the preceding section.
5.6. Other models

One important set of models that is mostly ignored the above considerations are classic probabilistic models (category [II.2.1]). For a description, see Section 4.1. Examples of combinatorial optimization algorithms employing such models are variants of EDA or ACO [6, 15]. Compared to the earlier presented works, these algorithms usually operate under different assumptions and are often applied to less expensive (although not necessarily cheap) problems. Still, they can be considered to use surrogate models. If necessary, a mapping between search and objective function can be established (cf. [104, 105]). Applications of ACO for (expensive) combinatorial problems, e.g., include the works by Cáreres et al. [81] and Teixeira et al. [69]. EDAs have been applied to problems in building design [106] and reactor fuel management [107].

From the strategy point of view, these distribution models can again be classified into the earlier described six categories. For instance, there are EDAs which employ a mapping approach (STR-4), e.g., EDAs for permutation problems using the random key mapping [108]. Other EDAs use problem-specific solutions (STR-2). One example are EDAs based on the (generalized) Mallows model for permutations [109, 110, 111]. These may also be categorized as being dissimilarity or distance-based (STR-6), but are not easily extended to arbitrary other distances.

Another set of models not explicitly mentioned yet are graph-based models like Bayesian networks or Markov random fields (MRFs). These are of interest, as they may directly encode the graph structure of the candidate solution representations [112]: Directed graphs in Bayesian networks and undirected graphs in MRFs. In this graph, discrete variables may be represented by nodes, and dependencies between variables by edges. Graph based models are closely linked to and have been employed in distribution based approaches, e.g., [113, 114, 104, 105, 106]. Mostly, these approaches can be classified as inherently discrete models (STR-3).

Kriging can be seen as a special case of Gaussian MRFs, that is, the fully connected graph case. However, Kriging does not model the graph structure of
the representation: nodes are candidate solutions, not variables.

6. Recent Trends and Developments

The model error is not the only criterion for selecting the surrogate. Recent approaches such as the evolvability learning of surrogates (EvoLS) algorithm implement local models for each offspring individually [115]. This results in an adaptive semi-partition [116] of the search space. In contrast to this local procedure, Bartz-Beielstein [1, 117] presents a global approach, which combines information from every surrogate, even from surrogates, which have a low accuracy. These approaches will be presented in Section 6.1 and 6.2, respectively.

6.1. Evolvability Learning of Surrogates

The naive approach of choosing the surrogate according to its estimated error on $f$ does not use the available information efficiently. Poorly selected experimental designs might introduce some bias: Models, which have a low estimated error in irrelevant regions of the search space can outperform models that have a higher error, but reflect the overall structure of the search space much better. This situation is illustrated in Figure 3. Additionally, Viana and Haftka [118] observe that an improved global accuracy of the surrogate is not necessary for obtaining the best result.

The EvoLS algorithm, which is introduced by Le et al. [115], extends the basic surrogate-assisted EA. The authors of EvoLS recommend selecting surrogate models that enhance search improvement in the context of optimization. EvoLS processes information about the (i) different fitness landscapes, (ii) state of the search, and (iii) characteristics of the search algorithm to statistically determine the so-called evolvability of each surrogate model. The evolvability of a surrogate model estimates the expected improvement of the objective function value that the new candidate solution, $\tilde{x}$, has gained after a local search has been performed on the related surrogate model. The local optimizer on the surrogate model guarantees (theoretically) convergence to the stationary point of the exact objective function [119], [120].
Figure 3: A simple, artificial example which illustrates that the surrogate model with the smallest prediction error is not always the best choice. The solid line represents the true cubic function $f$. A dash-dotted line is used to plot quadratic surrogate, whereas a dotted line depicts the linear surrogate. Consider a minimization problem. The initial design points are located in the interval $I_+ = \{ x : 0 \leq x \leq 1 \}$. Compared to the linear model, the quadratic surrogate has a relatively small prediction error in $I_+$. However, from a global search perspective, which also includes the grey shaded region, the quadratic model is misleading. The linear model has a larger prediction error in $I_+$, but will guide the search in the region of optimal values.

A minimalistic EvoLS algorithm is shown in Algorithm 4. The pseudo code presents a simplistic description of the fundamental states, e.g., the EvolutionaryVariation() operation in line 4 includes a parental selection, recombination, and mutation step.

The surrogate selection (line 6) uses the evolvability, which is a function of the candidate solution, $\bar{x}$, and the surrogate $M$. The model with the highest evolvability is used to perform a local search (line 7). The EvoLS algorithm presents a sophisticated way of exploiting local information during the surrogate modeling phase.

6.2. Sequential Parameter Optimization

Early versions of the sequential parameter optimization toolbox (SPOT) combined methods from design of experiments, response surface methodology, design and analysis of computer experiments, and regression trees for the analysis of algorithms [121, 122, 123]. The SPOT was developed as a toolbox for the analysis and for an understanding of the working principles of EAs. It implements a

1: \( t = 0. \) \( P(t) = \text{SetInitialPopulation}. \)
2: Evaluate\((P(t))\) on \( f. \)
3: while not TerminationCriterion() do
4: \( P'(t) = \text{EvolutionaryVariation}(P(t)). \)
5: for each \( \vec{x} \) in \( P'(t) \) do
6: \( M = \text{SurrogateSelection}(\vec{x}). \)
7: \( \vec{x}' = \text{LocalSearch}(M, \vec{x}). \)
8: Evaluate\((\vec{x}'). \) on \( f. \)
9: \( P(t + 1) = P(t) \cup \{\vec{x}'\}. \)
10: end for
11: \( t = t + 1. \)
12: end while

sequential, surrogate model based approach to optimization. Information from the surrogate model is used for analyzing the working principles of EAs. Effects of the the algorithm parameters, e.g., the population size or the type of recombination operator, and their interactions can be statistically analyzed. Nowadays, the SPOT is an established parameter tuner and an optimization algorithm, which has been extended in several ways. For example, Hutter et al. [124] benchmark an SPOT derivative, the so-called sequential model-based algorithm configuration (SMAC) procedure, with the black-box optimization benchmark (BBOB) set of test functions [125]. They demonstrate that with a small budget of \( 10 \times d \) evaluations of \( d \)-dimensional functions, SMAC in most cases outperforms the state-of-the-art blackbox optimizer covariance matrix adaptation evolution strategy (CMA-ES).

The most recent version, SPOT2, integrates a state-of-the-art ensemble learner. The SPOT2 ensemble engine can be briefly outlined as follows (see Algorithm 5): The portfolio or set of surrogate models, \( \mathfrak{M} \), includes a large amount of surrogates, such as regression trees and random forest, least angle regression (LARS), and Kriging, see line 2 in Algorithm 5. As such, it includes
several models that can also handle continuous, ordinal, categorical, and binary variables. Thus, SPOT2 can also deal with combinatorial optimization problems. The SPOT2 ensemble engine uses cross validation to select an improved model from the portfolio of candidate models [126]. It implements methods for creating a weighted combination of several surrogate models to build the improved model and methods, which use stacked generalization to combine several level-0 models of different types with one level-1 model into an ensemble [127]. The level-1 training algorithm is typically a relatively simple linear model, see line 5 in Algorithm 5. Figure 1 in [117] illustrates the stacking procedure.

Algorithm 5 Sequential parameter optimization with stacking
1: $t = 0$. \( P(t) = \text{SetInitialPopulation}(). \)
2: Select the set of cheap models \( \mathcal{R} \).
3: Evaluate\( (P(t)) \) on \( f \).
4: while not TerminationCriterion() do
5: Use \( P(t) \) to build a stacked model \( M(t) \) using \( \mathcal{R} \).
6: \( P'(t + 1) = \text{GlobalSearch}(M(t)) \).
7: Evaluate\( (P'(t + 1)) \) on \( f \).
8: \( P(t + 1) = P(t) \cup P'(t + 1) \).
9: \( t = t + 1 \).
10: end while

The implementation of the proposed stacking approach is available in the open source R package SPOT2, made available in the supplemental material of this article\(^2\).

Results indicate that the SPOT2 ensemble engine can lead to significant performance improvements, which is illustrated by the following example: Rebolledo et al. [128] present a comparison of different data driven modeling meth-

\(^2\)The SPOT2 software package and the supplemental material, which is necessary to run the following experiments, can be downloaded from the first author’s home page http://www.gm.fh-koeln.de/~bartz/SPOT/
Table 2: Comparison of the MSE values of individual models and a stacked ensemble. Smaller values are better. Best results are shown in bold face. Values are based on the gas sensor data example [128]. Numbers in brackets indicate standard deviations (based on ten repeats). Since Kriging and the linear model are deterministic, their standard deviation is zero. Numbers for the individual models differ from previous results [128], because different model implementations are used.

<table>
<thead>
<tr>
<th></th>
<th>Kriging</th>
<th>Linear Model</th>
<th>Random Forest</th>
<th>SPOT2 Ensemble</th>
</tr>
</thead>
<tbody>
<tr>
<td>sensor 1</td>
<td>0.495 (0)</td>
<td>0.593 (0)</td>
<td>0.594 (0.014)</td>
<td><strong>0.328 (0.085)</strong></td>
</tr>
<tr>
<td>sensor 2</td>
<td>0.419 (0)</td>
<td>0.577 (0)</td>
<td>0.463 (0.009)</td>
<td><strong>0.287 (0.066)</strong></td>
</tr>
</tbody>
</table>

ods. The models are built on industrial data for the development of a robust gas sensor. The data contains a limited amount of samples with high variance. The mean square error of the models implemented in a test dataset is used as the comparison strategy. Two sensors were tested in this comparison.

Table 2 compares the results received by a stacked ensemble with the results of individual models which are contained in the ensemble. Training and testing were repeated ten times, to account for randomness in the random forest model as well as the ensemble model. While Kriging is the best individual model, the ensemble consistently outperforms it. The larger variance in the ensemble results can be attributed to the dependence on the random sampling in the cross-validation procedure. In the level-1 model of the ensemble, Kriging received the highest coefficients. This real-world application example demonstrates the potential merit of stacked ensembles. The code of this example is made available in the supplemental material of this article.

7. Summary and Outlook

Model-based optimization (and in particular SBO) approaches are probably the most efficient methods for expensive and time-demanding real-world optimization problems. Although this observation is clearly true in the engineering domain, similar tendencies can be observed in many other domains, e.g., drug-design [129, 28].
Summarizing, the contribution of this article can be described as follows: A comprehensible taxonomy of MBO algorithms for global optimization problems (Figure 1) is proposed. A survey of SBO algorithms, which includes the most recent publications from continuous and the combinatorial problem domains, is given. Six strategies for dealing with modeling in combinatorial search spaces are developed. Working principles of two state-of-the-art SBO algorithms were shown: (i) EvoLS, which constructs a local metamodel for every new candidate solution, and (ii) SPOT2, which uses a global ensemble engine to combine a broad variety of surrogate models. The survey presented in the first sections of this article as well as the examples in Section 6 emphasize the trend to ensemble based metamodels. Due to the emerging-field nature of SBO, and especially combinatorial SBO, several challenges remain to be solved. This article concludes with a (subjective) selection of the most challenging problems.

7.1. Model Selection

The selection of an adequate surrogate plays a crucial role in the SBO process. The surrogate should reflect the underlying complex process as exact as possible and should be as simple as possible. The estimation of the model quality is an open research question. For example, the definition of the corresponding training sets for the holdout or cross-validation approaches represents a critical issue for the accuracy and efficiency of the surrogates. Viana and Haftka [118] report that it is beneficial to run EGO with multiple surrogates. Interestingly, they also observe that an improved global accuracy of the surrogate is not necessary for obtaining the best result.

The number of potential surrogate model types and selection strategies is huge. The earlier mentioned stacking approach (cf. Section 6.2) can prove to be a helpful starting point. Building stacked ensembles of surrogate models may help to understand how different approaches perform, and how they can interact. Besides providing more accurate predictions, stacking may enable researchers to design more promising modeling approaches for complex data structures.

Model selection is still a challenging research topic with many open questions,
which is especially true in the combinatorial domain. There, little guidelines are available, due to the relative scarcity of publications.

7.2. Definiteness

In case of similarity-based models (strategy STR-6, see Section 5.5), definiteness is a crucial issue. While first results are available for Kriging models [99], some important details require further attention. First, more extensive experiments are of interest. Secondly, more efficient handling of prediction for new data samples would be beneficial for performance in practice. And finally, a theoretical analysis could provide a more sound foundation for the existing approaches.

7.3. Dimensionality

The question of dimensionality, i.e., the number of variables, is an important issue. For continuous problems, it is often stated that distance-based models like Kriging perform poorly for larger dimensional problems. A rough threshold of approximately 20 dimensions is frequently specified for Kriging, e.g., see [7]. This is closely linked to the specific distance function [130]. For instance, Manhattan distance will be less affected by such issues, compared to Euclidean distance [130]. For the general, discrete case with an arbitrary distance function such knowledge is often not available. At the same time, feature selection or other dimensionality reduction methods may not be instantly available for problems with complex, discrete data representations. Thus, further research on dimensionality issues is recommendable. This is especially of interest for high dimensional real-world problems as, e.g., investigated by Smith et al. [71]. Simpson et al. [2] list some recent approaches to tackle this “curse of dimensionality” and problems related to generate adequate surrogates in high-dimensional and complex non-linear search spaces.

Dimensionality can also be interpreted as the number of samples, which leads to additional challenges. Nowadays, an increasing number of applications generate large data sets. Instead of very expensive and small data sets, huge data sets
have to be processed in this situation. Several applications from bioinformatics, social media, or climate science rely on the modeling of massive data sets with up to billions of data samples. Recently developed methods rely on special matrix structures (Kronecker or Toeplitz). For example, Wilson et al. [131] present early results for massively scalable Gaussian processes, which enables the use of Gaussian processes on billions of data points. The development of multi-fidelity surrogates, which use coarse grained models to obtain similar results as the exact model might be very interesting. Furthermore, methods for data sets that do not satisfy special matrix structures are of great interest.

7.4. Benchmarking

An important issue is the set of benchmark or test functions, used to evaluate algorithm and modeling performance. Previous approaches try to compose a set of test functions with many different features, e.g., by using step, linear, convex, and sinusoidal functions. A more recent approach [132] uses an infinite number of test problem instances to prevent an over fitting (or better: over learning) of the competing algorithms. The test problems are based on real-world problem instances, which are systematically modified.

While benchmarking is still not resolved for continuous model-based optimization, the situation is even less settled in the discrete domain. Of the few published, real-world, expensive, combinatorial problems, most are not openly accessible. Even in case of availability, the benchmark set would be rather small and the expense of computation would hinder broader experimental studies. Thus, most of the benchmark problems are well known cheap-to-compute problems, e.g., the traveling salesman problem (see the overview in Table 1). It is questionable whether performances estimated for these problems are actually representative for real-world expensive problems. In fact, the simple problem structure may be unable to give proper credit to complex models like Kriging. This may be one reason for results as, e.g., observed by Caceres et al. [81].
7.5. Multiple Objectives

Last, but not least, the problems discussed so far for single objective SBO exist also for model-based multi-objective optimization algorithms. The question of using global or local models is discussed in several publications, e.g., Isaacs et al. [133] present a local approach, which maintains an external archive, whereas Pilat and Neruda [134] present an approach by aggregating metamodels for evolutionary multiobjective and many-objective optimization. Horn et al. [135] present a taxonomy for model-based multi-objective optimization algorithms, which can be recommended as a starting point.

Acknowledgements

This work has been supported by the Bundesministerium für Wirtschaft und Energie under the grants KF3145101WM3 and KF3145103WM4. This work is part of a project that has received funding from the European Unions Horizon 2020 research and innovation program under grant agreement No 692286.

References


M. Kennedy, Predicting the output from a complex computer code when


[54] R. Jin, W. Chen, T. W. Simpson, Comparative studies of metamodeling techniques under multiple modelling criteria, Structural and


[76] M. Herrera, A. Guglielmetti, M. Xiao, R. Filomeno Coelho, Metamodel-assisted optimization based on multiple kernel regression for mixed vari-


[108] J. Ceberio, E. Irurozki, A. Mendiburu, J. A. Lozano, A review on estimation of distribution algorithms in permutation-based combinatorial opti-


Kontakt/Impressum

Diese Veröffentlichungen erscheinen im Rahmen der Schriftenreihe "CIplus". Alle Veröffentlichungen dieser Reihe können unter https://cos.bibl.th-koeln.de/home abgerufen werden.

Köln, Januar 2012

Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

Herausgeber / Editorship

Prof. Dr. Thomas Bartz-Beielstein,
Prof. Dr. Wolfgang Konen,
Prof. Dr. Boris Naujoks,
Prof. Dr. Horst Stenzel
Institute of Computer Science,
Faculty of Computer Science and Engineering Science,
TH Köln,
Steinmüllerallee 1,
51643 Gummersbach
url: www.ciplus-research.de

Schriftleitung und Ansprechpartner/ Contact editor’s office

Prof. Dr. Thomas Bartz-Beielstein,
Institute of Computer Science,
Faculty of Computer Science and Engineering Science,
TH Köln,
Steinmüllerallee 1, 51643 Gummersbach
phone: +49 2261 8196 6391
url: http://www.spotseven.de
eMail: thomas.bartz-beielstein@th-koeln.de

ISSN (online) 2194-2870
Supported by:

Federal Ministry for Economic Affairs and Energy

on the basis of a decision by the German Bundestag

Grant No. KF3145101WM3 and KF3145103WM4

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 692286.