Data mining for decision making in engineering optimal design

A. Mosavi

University of Debrecen, Faculty of Informatics, Hungary.

Received 02 March 2013; accepted 07 April 2013

*Corresponding author: a.mosavi@math.unideb.hu (A. Mosavi).

Abstract

Often in modeling the engineering optimization design problems, the value of objective function(s) is not clearly defined in terms of design variables. Instead it is obtained by some numerical analysis such as finite element structural analysis, fluid mechanics analysis, and thermodynamic analyses. Yet, the numerical analyses are considerably time consuming to obtain the final value of objective function(s). For the reason of reducing the number of analyses as few as possible, our methodology works as a supporting tool to the meta-models. The research in meta-modeling for multi-objective optimization are relatively young and there is still much research capacity to further explore. Here is shown that visualizing the problem on the basis of the randomly sampled geometrical big-data of computer aided design (CAD) and computer aided engineering (CAE) simulation results, combined with utilizing classification tool of data mining could be effective as a supporting system to the available meta-modeling approaches.

To evaluate the effectiveness of the proposed method, a case study in 3D wing optimal design is proposed. Discussion focusing on how effective the proposed methodology could be in further practical engineering design problems is presented.

Keywords: Data Mining, classification, Multi-objective Optimization, Engineering Optimization, Meta-Modeling.

1. Introduction

The research field of considering decision problems with multiple conflicting objectives is known as multiple criteria decision making (MCDM) [1]. Solving a multi-objective optimization problem has been characterized as supporting the decision maker (DM) in finding the best solution for the DM’s problem. DM and optimization typically create an interactive procedure for finding the most preferred solutions. Yet, despite the increasing level of complexity, it has been often tried to pay attention to improving all the defined objective functions instead of reducing or ignoring some of them. Although due to the increased complexity, this would apply complications where objective functions are visualized by trade-off analysis methods as well studied in [9, 10, 25, 26, 35, 37].

According to [1], the general form of the multi-objective optimization problems can be stated as;

Minimize \( \mathbf{f}(\mathbf{x}) = \{f_1(\mathbf{x}), ..., f_m(\mathbf{x})\} \), Subjected to \( \mathbf{x} \in \Omega \), where \( \mathbf{x} \in \mathbb{R}^n \) is a vector of \( n \) decision variables; \( \mathbf{x} \subset \mathbb{R}^n \) is the feasible region and is specified as a set of constraints on the decision variables; \( \mathbf{f} : \Omega \rightarrow \mathbb{R}^m \) is made of objective functions subjected to be minimization. Objective vectors are images of decision vectors written as \( \mathbf{z} = \mathbf{f}(\mathbf{x}) = \{f_1(\mathbf{x}), ..., f_m(\mathbf{x})\} \). Yet an objective vector is considered optimal if none of its components can be improved without worsening at least one of the others. An objective vector \( \mathbf{z} \) is said to dominate \( \mathbf{z}' \), denoted as \( \mathbf{z} < \mathbf{z}' \), if \( z_k \leq z'_k \) for all \( k \) and there exists at least one \( h \) that \( z_h \leq z'_h \). A point \( \mathbf{x} \) is Pareto optimal if there is no other \( \mathbf{x} \in \Omega \) such that \( \mathbf{f}(\mathbf{x}) \) dominates \( \mathbf{f}(\hat{\mathbf{x}}) \). The set of Pareto optimal points is called Pareto set (PS). And the corresponding set of Pareto optimal objective vectors is called Pareto front (PF).
Solving a multi-objective optimization problem would be done by providing the DM with the optimal solution according to some certain utility criteria allowing to choose among competing PF. Such utility criteria are often inconsistent, difficult to formalize and subjected to revision. The complete process of MCDM has two parts (1) multi-objective optimization process which tries to find the PF solutions (2) decision making process which tries to make the best decision out of the possible choices. In dealing with increased complexity, this paper focuses on the first part which mostly deals with variables, constraints and objective functions.

1.1. Computational intelligence and multi-objective optimization

Developing the methods for multi-objective optimization using computational intelligence along with real applications appeared to be quite young. However it has been observed that techniques of computational intelligence are indeed effective [3, 7, 15, 27]. On the other hand, the techniques of multi-objective optimization by themselves can also be applied to develop and to improve the effective methods in computational intelligence [2]. Currently there are many computational intelligence-based algorithms available to generate PF [1, 16, 17, 30]. However, it is still difficult to generate and visualize the PF in the cases with more than three objectives. In this situation, methods of sequential approximate optimization of computational intelligence with meta-modeling are recognized to be very effective in a series of practical problems [1, 4].

1.2 Meta-modeling and multi-objective optimization in shape optimization

Meta-modeling is a method for building simple and computationally inexpensive models, which replicate the complex relationships. However the research in meta-modeling for multi-objective optimization is relatively young and there is still much to do. So far there existed only a few standards for comparisons of methods, and little is yet known about the relative performance and effectiveness of different approaches [4, 15]. The most famous methods of Meta-modeling are known as response surface methods (RSM) and design of experiments (DOE). Although it is concluded in previous studies [16, 18, 19, 20], in the future research, scalability of MCDM models in terms of variables’ dimension and objective space’s dimension will become more demanding. This is because the models have to be capable of dealing with higher computation cost, noise and uncertainties.

According to [18], the application of meta-modeling optimization methods in industrial optimization problems is discussed. Some of the major difficulties in real-life engineering design problems counted: (1) there are numerous objective functions to be involved, (2) the function form of criteria is a black box, which cannot be explicitly given in terms of design variables, and (3) there are a huge number of unranked and non-organized input variables to be considered. Additionally in engineering design problems, often the value of objective functions is not clearly defined in terms of design variables. Instead it is obtained by some numerical analyses, such as FE structural analysis [34, 37], fluid mechanics analysis [7, 16, 17, 32], thermodynamic analysis [30], chemical reactions [3]. These analyses for obtaining a single value for an objective function are often time consuming. Considering the high computation costs, the number of CAE evaluations/calculations are subjected to minimization with the aid of meta-models [18]. In order to make the number of analyses as few as possible, sequential approximate optimization is one of the possible methods, utilizing machine learning techniques for identifying the form of objective functions and optimizing the predicted objective function. Machine learning techniques have been applied for approximating the black-box of CAE function in many practical projects [1, 9, 10, 25, 37]. Although the major problems in these realms would be (1) how to approach an ideal approximation of the objective function based on as few sample data as possible (2) how to choose additional data effectively. The objective functions are modeled by fitting a function through the evaluated points. This model is then used to help the prediction value of future search points. Therefore, those high performance regions of design space can be identified more rapidly. Moreover the aspects of dimensionality, noise and expensiveness of evaluations are related to method selection [32]. However, according to Bruyneel et al. [18] for the multi-objective capable version of meta-modeling algorithms further aspects such as the improvement in a Pareto approximation set and modeling the objective function must be considered.

Today, numerical methods make it possible to obtain models or simulations of quite complex and large scale systems [7, 8, 20, 22]. But there are still difficulties when the system is being modeled numerically. In this situation, modeling the
simplified models is an effective method, generating a simple model that captures only the relevant input and output variables instead of modeling the whole design space [3, 20, 22]. The increasing desire to apply optimization methods in expensive CAE domains is driving forward the research in meta-modeling. The RSM is probably the most widely applied to meta-modeling. The process of a meta-model from big data is related to classical regression methods and also to machine learning [4, 37]. When the model is updated using new samples, classical DOE principles are not effective. In meta-modeling, the training data sets are often highly correlated, which can affect the estimation of goodness of fit and generalization performance. Yet Meta-modeling brings together a number of different fields to tackle the problem on optimizing the expensive functions. On the other hand the classical DOE methods with employing evolutionary algorithms have delivered more advantages in this realm. Figure 1 describes the common arrangement of meta-modeling tools in multi-objective optimization processes of engineering design. It is worth mentioning that the other well-known CAD-Optimization integrations for shape optimization e.g. [24, 29, 31] would also follow the described scheme.

![Figure 1. Meta-modeling tools in multi-objective optimization process.](image)

2. Data mining classification in engineering design applications

The particular advantage of evolutionary algorithms (EAs) [11] in the multi-objective optimization (EMO) applications [19] is that they work with a population of solutions. Therefore, they can search for several Pareto optimal solutions providing the DM with a set of alternatives to choose from [14]. EMO-based techniques have an application where mathematical-based methods have difficulties with. EMO are also helpful in knowledge discovery related tasks in particular for mining the data samples achieved from CAE and CAD systems [29, 31]. Useful mined information from the obtained EMO trade-off solutions have been considered in many real-life engineering design problems.

2.1. Classifications

Finding useful information in large volumes of data drives the development of data mining procedure forward. Data mining classification process refers to the induction of rules that discriminate between organized data in several classes so as to gain predictive power [5]. There are some example applications of data mining classification in evolutionary multi-objective optimization available in the literature of [1, 6, 12, 19] where the goal of the classification algorithms is to discover rules by accessing the training sets. Then the discovered rules are evaluated using the test sets, which could not be seen during the training tasks [5].

In the classification procedures, the main goal is to use observed data to build a model, which is able to predict the categorical or nominal class of a dependent variable given the value of the independent variables [5]. Obayashi [12] for the reason of mining the engineering multi-objective optimization and visualization data applied self-organizing maps (SOM) along with a data clustering method. Moreover Witkowski et al. [13] and Mosavi [7, 20, 22] used classification tools of data mining for decision making supporting process to multi-objective optimization.

2.2. Modeling the problem

According to [1], before any optimization takes place, the problem must first be accurately modeled. In this case, identifying all the dimensions of the problem, such as formulation of the optimization problem with specifying decision variables, objectives, constraints, and variable bounds is an important task. Here the methodology proposes that mining the available sample data before actual modeling will indeed help to better model the problem as it delivers more information about the importance of input variables and could in fact rank the input variables. The proposed method of classification, also earlier utilized in [7, 20, 22], presented in Figure 2, is set to mine the input variables which are in fact associated with the final CAE data.
3. Study case; three-objective and 42-variale optimization problem

The applications in engineering optimal design have numerous disciplines to bring into the consideration. In mechanical engineering, the structural simulation is tightly integrated more than one discipline [18, 21, 22, 23, 24, 25]. Meanwhile, the trend nowadays is to utilize independent computational codes for each discipline [32]. In this situation, the aim of MCDM tools is to develop methods in order to guarantee that all physical variables are involved in the model. Bo et al. [28] in aerodynamic optimization of a 3D wing has tried to utilize the multi-objective optimization techniques in a multidisciplinary environment.

In the similar cases [20, 24, 29, 32] in order to approach the optimal shape in an aerospace engineering optimization problem, the multi-objective optimization techniques are necessary to deal with all important objectives and variables efficiently. Here the optimization challenge is to identify as many optimal designs as possible to provide a choice of better decision. However with an increased number of design variables the modeling task, in a multidisciplinary environment, is getting even more complicated. Therefore the multi-objective optimization tasks become more difficult with the increasing number of variables [20, 35]. Although the recent advances in parametric CAD/CAE integrations [24, 29, 31] have reduced the complexity of the approach in some levels.

The airfoil of Figure 3 part (a) is subjected for shape improvement. The shape needs to be optimized in order to deliver minimum displacement distribution in terms of applied pressure on the surface. Figure 3, part (b) shows the basic curves of the surface modeled by S-plines. Here the proposed S-pline geometrical modeling methodology of Albers et al. [36] is successfully adapted and utilized. In the study case for modeling the 3D wing surface, four curve profiles have been with 42 points utilized. The coordinates of all points are supplied by a digitizer in which each point includes three dimensions of X, Y, and Z. Consequently the case, by adding the variable constraints, would include 126 columns plus three objectives which are going to highly increase the complexity. In fact, an optimal configuration of 42 variables supposed to satisfy the following three described objectives.

The objectives are listed as follows:

- **Objective 1:** Minimizing the displacement distribution in the airfoil for constant pressure value of $\alpha$.
- **Objective 2:** Minimizing the displacement distribution in the airfoil for constant pressure value of $2\alpha$.
- **Objective 3:** Minimizing the displacement distribution in the airfoil in constant pressure value of $4\alpha$.

In the described multi-objective optimization problem the number of variables is subjected to minimization before the multi-objective optimization modeling process takes place in order to evolve a large scale design space to the smaller and much more handy design space. Here the proposed and utilized model reduction methodology differs from the previous study Filomeno et al. [35] in terms of applicability and ease of use in general multi-objective optimization design applications.
The dataset of big data for data mining is supplied from the Table I. The table has gathered a collection of initial dataset including shapes' geometries and simulation results from five CAE calculations, based on random initial values of variables, which in the proposed method will be mined. In the next section, the discussion of how the dataset of five random CAE calculations are being utilized for creating the smaller design space for a multi-objective optimization model is made.

4. Methodology and experimental results
The effectiveness of data mining tools in multi-objective optimization problems presented by Coello et al. [2] and earlier in [5] the classification rules for evolutionary multi-objective algorithms were well implemented, in which along with the research work of Witkowski et al. [13] forms the proposed methodology working via a novel workflow. The workflow of data mining procedure methodology is described in Figure 4. In this method, the classification task is utilized to create several classifiers or decision trees. In the next steps, the most important variables, which have more effects on the objectives, are detected.

Table I. Training dataset including five CAE calculations' results

<table>
<thead>
<tr>
<th>Variables Configuration</th>
<th>CAD Model</th>
<th>Displacement Distribution</th>
<th>Objective Results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>O1≈×  O2≈×  O3≈×</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td>O1≈×  O2≈×  O3≈×</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>O1≈×  O2≈×  O3≈×</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>O1≈×  O2≈×  O3≈×</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>O1≈×  O2≈×  O3≈×</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>O1≈×  O2≈×  O3≈×</td>
</tr>
</tbody>
</table>

4.1 Methodology and experimental results

The effectiveness of data mining tools in multi-objective optimization problems presented by Coello et al. [2] and earlier in [5] the classification rules for evolutionary multi-objective algorithms were well implemented, in which along with the research work of Witkowski et al. [13] forms the proposed methodology working via a novel workflow. The workflow of data mining procedure methodology is described in Figure 4. In this method, the classification task is utilized to create several classifiers or decision trees. In the next steps, the most important variables, which have more effects on the objectives, are detected.

Regressions and model trees are constructed by a decision tree building an initial tree. However, most decision tree algorithms choose the splitting attribute to maximize the information gain. It is appropriate for numeric prediction to minimize the intra subset variation in the class values under each branch.

The splitting criterion is used to determine which variable is better to split the portion T of the training data. Based on the treating the standard deviation of the objective values in T as a measure of the error and calculation the expected reduction in error as a result of testing each variable is calculated. Meanwhile the variables, which maximize the expected error reduction, are chosen for splitting. The splitting process terminates when the objective values of the instances vary very slightly, that is, when their standard deviation has only a small fraction of the standard deviation of the original instance set. Splitting also terminates when just a few instances remain. Experiments show that the obtained results are not very sensitive to the exact choice of these thresholds. Data mining classifier package of Weka provides implementations of learning algorithms and dataset which could be preprocessed and fed into a learning scheme, and analyze the resulting classifier and its performance. The workbench includes methods for all the standard data mining problems such as regression, classification, clustering, association rule mining, and attribute selection. Weka also includes many data visualization facilities and data preprocessing tools. Here three different data mining classification algorithms i.e. J48, BFTree, LADTree are applied and their performance is compared to choose attribute importance. The mean absolute error (MAE) and root mean squared error (RMSE) of the class probability is estimated and assigned by the algorithm output. The RMSE is the square root of the average quadratic loss and the MAE is calculated in a similar way using the absolute instead of the squared difference.

The comparison between importance ranking results is obtained by our experiments listed in Table II. It is concluded that in the worst case, more than 55% variable reduction is achieved. As one can see, BFTree and J48 algorithms have classified the datasets with less number of variables. While in LADTree algorithms, at least seven variables have
utilized to classify dataset. The variables number 15 and 24 play much more important role in effecting the first objective ($O_1$). Variables number 41 and 35 also have the more effects on third objective ($O_3$) as well. According to the experimental results, it is possible to optimize the model by reducing the 45% number of variables. In Table II, two types of classification error (MAE, RMSE) are shown for all algorithms corresponding to different class of objectives.

Table 2. Variables importance ranking for three classification methods

<table>
<thead>
<tr>
<th>Classification Method</th>
<th>MAE</th>
<th>RMSE</th>
<th>Effective Variables</th>
<th>Objective(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BFTree</td>
<td>0.370</td>
<td>0.517</td>
<td>15, 23, 41</td>
<td>$O_1$, $O_2$, $O_3$</td>
</tr>
<tr>
<td></td>
<td>0.412</td>
<td>0.519</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.418</td>
<td>0.555</td>
<td></td>
<td></td>
</tr>
<tr>
<td>J48</td>
<td>0.309</td>
<td>0.514</td>
<td>15, 24, 13, 35, 41</td>
<td>$O_1$, $O_2$, $O_3$</td>
</tr>
<tr>
<td></td>
<td>0.498</td>
<td>0.642</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.338</td>
<td>0.590</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LAD Tree</td>
<td>0.277</td>
<td>0.500</td>
<td>15, 24, 23, 32, 41, 39, 22, 18, 15, 42, 2, 17</td>
<td>$O_1$, $O_2$, $O_3$</td>
</tr>
<tr>
<td></td>
<td>0.604</td>
<td>0.769</td>
<td>20, 41, 35, 9, 17, 11, 38, 37, 16</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.365</td>
<td>0.584</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5. Conclusions
In order to extract more information from the optimization variables in a reasonable way, the classification task of data mining has been applied. Variables were ranked and organized utilizing three different classification algorithms. The results show the reduced number of variables speeds up and scales up the process of optimization within a preprocessing step. The utilized data mining tool has found to be effective in this regard. Additionally, it is evidenced that the growing complexity can be handled by a preprocessing step utilizing data mining classification tools. The modified methodology is demonstrated successfully in the framework and the author believes that the process is simple and fast. Future research should focus on the effectiveness of the proposed data reduction process. Also, trying other data mining tasks such as clustering, association rules, and comparison could be beneficial. Although in real-life applications where the optimal design problem has to be considered by inclusion of multiple criteria, a combination of the proposed method with the other developed MCDM tools [38-46] would be effective.

References


and International Mode Frontier users’ Meeting, Trieste, 2010.


[38] Mosavi, “A MCDM Software Tool for the Automated Design Environments,” 26th European Conference on Operational Research, Rome 2013, EURO - INFORMS XXVI.


