

21st European Symposium on Computer Aided Process Engineering – ESCAPE 21
E.N. Pistikopoulos, M.C. Georgiadis and A.C. Kokossis (Editors)
© 2011 Elsevier B.V. All rights reserved..

A data mining approach for efficient systems optimization under uncertainty using stochastic search methods

Garyfallos Giannakoudis^a, Athanasios I. Papadopoulos^a, Panos Seferlis^{a,b},
Spyros Voutetakis^a

^a*Chemical Process Engineering Research Institute, Centre for Research and
Technology-Hellas, 6th km Harilaou Thermi Road, 57001, Thessaloniki, Greece*

^b*Department of Mechanical Engineering, Aristotle University of Thessaloniki, 54124,
Thessaloniki, Greece*

Abstract

This work presents a novel approach for efficient systems design under uncertainty that uses data mining and model fitting methods during optimization to significantly reduce the associated computational effort. The proposed approach is implemented as part of a modified stochastic annealing algorithm, but remains independent of the employed optimization method. A numerical example and a case study on a stand-alone system for power generation from renewable energy sources are used to illustrate the merits of the developments. The obtained results indicate robustness and efficiency in terms of solution quality and computational performance, respectively.

Keywords: Systems optimization, uncertainty, data mining, Stochastic Annealing, renewable energy.

1. Introduction

Stochastic search methods such as Stochastic Annealing (StA) and Stochastic Genetic Algorithms (SGA) [1-4] have been proposed in recent years to address the optimization under uncertainty of process systems. The underlying algorithmic philosophy employed to treat uncertainty involves the use of probability distributions to generate samples which are introduced individually into the simulation of system models. This enables the emulation of effects caused by the uncertain parameters in the addressed optimization problem. Apparently, the number of utilized samples is of crucial importance. Large numbers of samples are required to maintain a realistic representation of the uncertain parameter distribution but at the expense of reduced computational efficiency. This is due to the increased computational effort required to simulate the effects of each sample through the employed system model during optimization. This major issue has been previously addressed [2, 3] by employing efficient sampling techniques and strategies that allow a variable sampling schedule throughout the optimization procedure. Fewer samples are allowed at initial optimization iterations, which are then increased significantly as the algorithm gradually proceeds to

termination. However, their utilization often requires significant computational effort as the random selection of a large number of samples is not prevented, even at initial optimization iterations. Furthermore, high numbers of samples towards termination still result to an increased computational burden for large-scale problems involving detailed system models and combinatorial complexities.

2. Proposed method

This work proposes the combined use of data mining and model fitting in the course of optimization to enable efficient management of the sampling procedure, employed to treat the considered uncertain parameters. Figure 1 illustrates the proposed approach as an extensive modification to the StA algorithm [1-3]. The novel algorithmic sequence is highlighted within the dashed frame. The implemented modifications are independent of the employed optimization algorithm, as they do not intervene with decision making operations that are distinctive of particular algorithms. While Hammersley sampling is employed in this work, any other sampling technique can also be utilized.

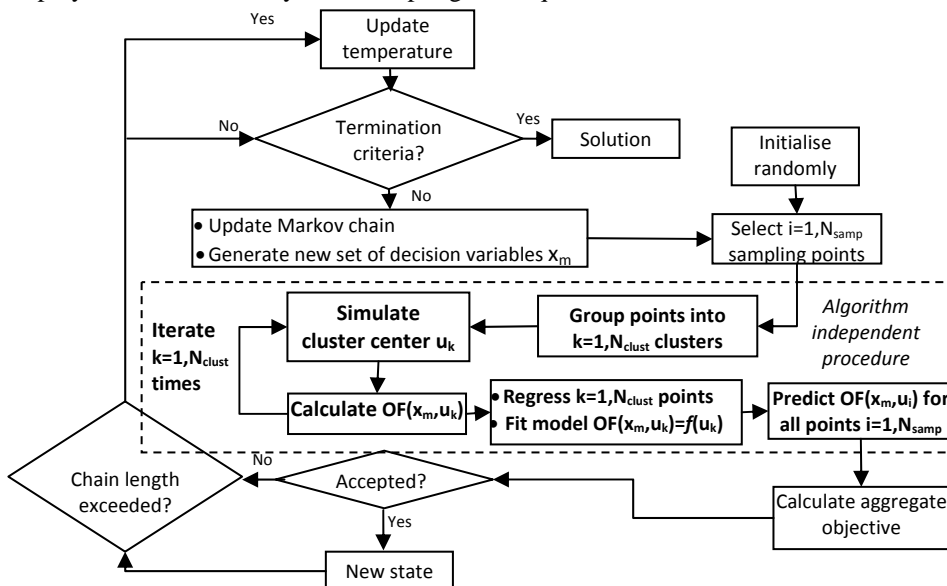


Figure 1: Proposed data mining method as part of a modified Stochastic Annealing algorithm

2.1. Description

Initially, a clustering method is used to generate $k=1, N_{clust}$ coherent groups (clusters) of similar points out of the entire set of the selected sampling points $i=1, N_{samp}$, used for the representation of the uncertain parameters vector (u). Statistical cluster centers (u_k) are then calculated for each group, which lie in close proximity to the entire data contained in each cluster. As a result, each cluster center can be considered a valid representative for all the data (sampling points) contained in the cluster. Subsequently all cluster centers, instead of all available sampling points, are introduced to simulations using a system model to calculate the objective function value $OF(x_m, u_k)$ that corresponds to

each center (u_k) (where x_m represents the vector of decision variables). In this respect, the available cluster center points (independent parameters) are then used in conjunction with their corresponding objective function values (dependent parameters) to calculate the regression coefficients of a continuous model. This model represents the employed $OF(x_m, u_k)$ as a mathematical function of the cluster centers [$OF(x_m, u_k) = f(u_k)$]. The objective function values $OF(x_m, u_i)$ that correspond to the remaining sampling points (u_i), contained in each cluster, can now be calculated using the developed predictive model, hence avoiding the time consuming simulations based on the system model.

2.2. Implementation details

The proposed approach enables the use of constantly large numbers of sampling points regardless of the size of the optimization problem addressed or the stage of the performed optimization search. The number of generated clusters is an important parameter that affects the performance of the method. A large number of clusters results to fewer points within each cluster. This enables an improved representation from the derived center of all the cluster points and results to accurate predictions from the regression model. However, increasing the number of clusters also results to further time-consuming simulations. An automated statistical method is used [5] to maintain the number of clusters considerably lower compared to the sampled set of uncertain parameter values, while facilitating accurate model predictions.

The fitted model provides objective function value predictions that are either identical or lie within very close proximity to the values calculated through simulations. This is verified by use of the R^2 coefficient of multiple determination, which is calculated in three steps. Firstly, predictions are obtained through the regression model for $OF(x_m, u_i)$ values. Subsequently, the predictions are used to replace their corresponding sampling points (u_i) that exist within each one of the original clusters. Finally, a new cluster center is derived for each cluster based on the objective function values (and not the sampled points as previously). This center represents the predicted objective function values that lie within each cluster. If it is similar to the objective function values obtained through model simulations for each corresponding cluster center, then the regression model provides accurate predictions. This similarity is measured through R^2 . The number of regression terms employed in the model is derived through statistical F-tests for model adequacy, also used to evaluate the correctness of R^2 .

2.3. Numerical example

The proposed method is illustrated through a numerical example that employs the following cost model (details available in [1]):

$$OF_1(y_1, y_2, y_3, u_1, u_2) = \sum_{i=1}^{y_1} ((y_1 - 3)^2 + (u_1 y_2^i - 3)^2 + (u_2 y_3^i - 3)^2) \quad (1)$$

Terms y_1, y_2, y_3 represent the decision variables. The uncertain parameters u_1 and u_2 follow the probability distributions shown in Table 1, which also shows the clustering ranges considered and the employed regression model. The regression coefficients α_i ($i=1,6$) are recalculated in each algorithmic iteration. In all cases the performance of the StACMF algorithm (StA with clustering and model fitting) is compared with an

adaptation of StA developed in this work. Their comparative performance is measured based on the ratio of the number of simulations performed by the two algorithms ($N_{\text{StACMF}}/N_{\text{StA}}$) to achieve optimality. The number of allowed samples is constantly 150 for StACMF, while sampling for StA is allowed to vary in the range [20, 150].

Table 1: Data and optimization-computational performance results for the numerical example

Case	u_1	u_2	Clustering range	Optimum values for $(y_1), (y_2), (y_3)$	R^2	Performance ratio
1	N(0,2)	N(0,2)	25-35	(3),(3,3,3),(3,3,3)	>0.999	0.39
2	N(0,2)	N(0,2)	15-25	(3),(3,3,3),(3,3,3)	>0.999	0.26
3	N(0,2)	U(1.5,3)	20-30	(3),(3,3,3),(1,1,1)	>0.996	0.28
Regression model: $OF(u_1, u_2) = a_1 + a_2 u_1 + a_3 u_2 + a_4 u_1 u_2 + a_5 u_2^2 + a_6 u_1 u_2^2$						

In all three cases the two algorithms found the same optimum solution. The obtained results indicate that StACMF is significantly faster, as the number of required simulations is only a small fraction of those required by StA. The value of R^2 is very high in all cases, indicating that the employed model provides accurate predictions. The minor inaccuracies in the predictions ($R^2 < 1$) do not prohibit the identification of the optimum solution by StACMF. The use of a lower clustering range (fewer clusters) in case 2 results to improved performance compared to case 1, while the optimum solution is still obtained. The simultaneous use of different distributions in case 3 does not affect the optimization and computational performance of StACMF compared to StA.

3. Case study

3.1. Background

The proposed approach is applied to the design optimization of a hybrid system for power generation from renewable energy sources, with medium- to long-term energy storage capabilities in the form of hydrogen. It consists of photovoltaic panels, wind generators, chemical accumulators, an electrolyser, a fuel cell, a compressor, hydrogen storage tanks and a diesel generator. Details can be found in [6]. The design of such a system involves increased uncertainty due to unpredictable weather variability and equipment efficiency changes. The optimization aims to minimize the net present value (NPV) of investment for a ten year operating period. The considered decision variables are 8, namely the number of PV panels (n_{pv}), the number of the wind generators (n_{wg}), the nominal capacity of the accumulators (n_{acc}), the maximum operating power of the electrolyzer ($P_{\text{max,e}}$), the capacity of the intermediate (V_b) hydrogen storage tanks, the nominal power of the fuel cell ($P_{\text{op,fc}}$) and the upper (SOC_{max}) and lower (SOC_{min}) limits of the stage of charge of the accumulators. The considered uncertain parameters are 4 and involve the solar radiation (u_1) and wind speed (u_2) as well as the efficiencies of the electrolyzer (u_3) and of the fuel cell (u_4). The number of allowed samples for the two algorithms is similar to that used in the numerical example, whereas the range of allowed clusters is [25, 35].

3.2. Results and discussion

The optimum OF value identified by StA is slightly better than StACMF (Table 2). This is a reasonably small deviation considering the high combinatorial complexity of the

A data mining approach for efficient systems optimization under uncertainty using stochastic search methods

315

problem. It is also characteristic of stochastic search methods that often converge to a narrow distribution of similar solutions. The small difference corresponds to slightly fewer PV panels identified in StA and slightly different SOC limits (up to 4%), while the optimum values for the remaining decision parameters are identical. Even though the design problem involves multiple decision variables and uncertain parameters, the regression model consists of only eight terms. Only uncertain parameters u_1 and u_2 are necessary in the model, based on the statistical tests performed for the determination of the required terms which implies that only u_1 and u_2 have significant effect on the objective function value under the operating conditions assumed in the case. Design of the system under different weather data (e.g., another location) would have resulted in fitting models with significant contribution by the equipment efficiencies u_3 and u_4 . In the vast majority of the optimization iterations R^2 is maintained over 0.98. This indicates reasonably good fitting using a simple model, compared to the much more complex system models that are avoided. The time performance per iteration is almost three times better with StACMF, indicating significant gains for time-consuming design problems. Such benefits are combined with the constantly used 150 samples throughout the optimization in StACMF.

Table 2: Data and performance results of StA and StACMF algorithms in case study.

Method	Average number of simulations per iteration	Average CPU time (sec) per simulation	Average CPU time (sec) for clustering + model fitting calculations per iteration	Average total CPU time (sec) per iteration	OF [k€]
StA	79	0.0642	-	5.0726	-46.205
StACMF	28	0.0642	0.0143	1.8122	-46.268
Regression model					
$OF(u_1, u_2) = b_o + b_1 \cdot u_1 + b_2 \cdot u_2 + b_3 \cdot u_1^2 + b_4 \cdot u_2^2 + b_5 \cdot u_1 \cdot u_2 + b_6 \cdot u_2 \cdot u_1^2 + b_7 \cdot u_1 \cdot u_2^2$					

4. Concluding remarks

The proposed approach increases the computational efficiency in systems optimization problems under uncertainty, while constantly maintaining an inclusive representation of the uncertain parameters throughout the optimization search. The implementation of clustering and model fitting are fast and computationally insignificant compared to numerous system model simulations required in existing StA implementations. The approach can handle increased numbers of decision variables and uncertain parameters without making the optimization computational effort prohibitive.

References

- [1] Painton L., Diwekar U., (1995), *Europ. J. Oper. Res.*, 83, 489-502.
- [2] Chaudhuri, P. Diwekar U., (1996), *AICHE J.*, 42, 3, 742-752.
- [3] Chaudhuri P., Diwekar U., (1999), *AICHE J.*, 45, 8, 1671-1687.
- [4] Xu W., Diwekar U. (2005), *Ind. Eng. Chem. Res.*, 44, 7132-7137.
- [5] Papadopoulos A.I., Linke P., (2006), *Chem. Eng. Sci.*, 61(19), 6316-6336.
- [6] Giannakoudis G., Papadopoulos A.I., Seferlis P., Vouetakis P., (2010), *Int. J. Hyd. Ener.*, 35, 872-891.