Computers and Chemical Engineering 108 (2018) 276–288

Contents lists available at ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/compchemeng

Evaluating smart sampling for constructing multidimensional surrogate models

Sushant S. Garud [a], Iftekhar A. Karimi [a,*], George P.E. Brownbridge [b], Markus Kraft [c,d]

[a] Department of Chemical & Biomolecular Engineering, National University of Singapore, 4 Engineering Drive 4, Singapore 117576, Singapore
[b] CMCL Innovations, Sheraton House, Castle Park, Cambridge CB3 0AX, United Kingdom
[c] Department of Chemical Engineering & Biotechnology, University of Cambridge, West Cambridge Site, Philippa Fawcett Drive, Cambridge CB3 0AS, UK
[d] School of Chemical & Biomedical Engineering, Nanyang Technological University, 62 Nanyang Drive, Singapore 637459, Singapore

ARTICLE INFO

Article history:
Received 22 June 2017
Received in revised form 25 August 2017
Accepted 19 September 2017
Available online 22 September 2017

Keywords:
Adaptive sampling
Experimental design
Surrogate model
Process flowsheet

ABSTRACT

In this article, we extensively evaluate the smart sampling algorithm (SSA) developed by Garud et al. (2017a) for constructing multidimensional surrogate models. Our numerical evaluation shows that SSA outperforms Sobol sampling (QS) for polynomial and kriging surrogates on a diverse test bed of 13 functions. Furthermore, we compare the robustness of SSA against QS by evaluating them over ranges of domain dimensions and edge length/s. SSA shows consistently better performance than QS making it viable for a broad spectrum of applications. Besides this, we show that SSA performs very well compared to the existing adaptive techniques, especially for the high dimensional case. Finally, we demonstrate the practicality of SSA by employing it for three case studies. Overall, SSA is a promising approach for constructing multidimensional surrogates at significantly reduced computational cost.

1. Introduction

Process simulators are commonly used to model, study, and analyze complex nonlinear physicochemical systems. However, such simulations are generally computationally intensive, thus, prohibiting their repeated evaluations in a typical analysis procedure. Moreover, the custom-made process simulators are often black-box in nature. Hence, no system information is available to the users without evaluating an instance of this costly simulation. On these accounts, it is beneficial to convert such high-fidelity simulations into computationally inexpensive surrogate models that capture essential features with reasonable numerical accuracy. Surrogate modeling, also known as metamodeling or response surface model, is a technique to generate a mathematical or numerical representation of a complex system based on some sampled input-output data. In a philosophical discussion on the future of computational modeling, Kraft and Mosbach (2010) highlight the importance of approximation techniques and experimental designs (sampling techniques) in tackling complex multi-scale systems. The quality of any surrogate approximation depends on a sampling technique used to generate the input-output data and a surrogate modeling technique used to build the approximation. The literature (Shan and Wang, 2010) has several forms of surrogate models like polynomial response surface model (PRSM), high dimensional model representation (HDMR), kriging, radial basis functions (RBFs), support vector regression (SVR), artificial neural networks (ANNs), etc. Furthermore, many works (Henao and Maravelias, 2011, 2010; Caballero and Grossmann, 2008) have employed these techniques in the context of various physicochemical systems. Nonetheless, the current work focuses on the critical evaluation of a smart and adaptive sampling approach for multidimensional surrogate construction paradigms.

Commonly used sampling techniques employ uniform, quasi-random, or systematic distributions (Pronzato and Müller, 2012; Koehler and Owen, 1996). Examples are factorial design or grid sampling, random sampling, Latin hypercube sampling, orthogonal arrays, Hammersley points, Sobol sampling (QS), etc. A recent review by Garud et al. (2017b) classifies the literature on sampling techniques into three major categories viz. static system-free, static system-aided, and adaptive-hybrid. It discusses each of them thoroughly and identifies their advantages and disadvantages. The static techniques are often prone to the curse of dimensionality. Moreover, they can result in under/oversampling and thus, resulting in poor system approximation (Garud et al., 2017a). In order to tackle these issues, a new upcoming class of modern DoE (design of experiments) called adaptive sampling (sequential sampling) has gained attention from the research community over the past few years. Adaptive sampling approach has two vital advantages

* Corresponding author.
E-mail address: cheiak@nus.edu.sg (I.A. Karimi).

http://dx.doi.org/10.1016/j.compchemeng.2017.09.016
0098-1354/

Abbreviations

Abbreviations

ANN	artificial neural network
CC	clustering constraint
CCU	carbon capture unit
CDM	crowding distance metric
CSTR	continuously stirred tank reactor
CV	cross validation
CVE	cross validation error
DEA	diethanolamine
DF	departure function
DoE	design of experiments
DT	Delaunay triangulation
EE	expected error
HDMR	high dimensional model representation
HM-CI	Hessian matrix based curvature information
JK	Jackknifing
LOLA	local linear approximation
MD	Mahalanobis distance
ME	maximum entropy
Mm	maximin distance
MoDS	model development suite
MSD	maximum scaled distance
MSE	maximum sampling error
NLP	nonlinear programming problem
NN	nearest neighbor
PE	pooled error
PRSM	polynomial response surface model
QS	Sobol sampling
RBF	radial basis function
RMSE	root mean squared error
SSA	smart sampling algorithm
SVR	support vector regression
VT	Voronoi tessellation
WCE	weighted cumulative error

Notation: Subscripts

b	index for the basis functions in kriging
m	index for elements of response/output variables' vector
n	index for elements of design/input variables' vector

Superscripts

i	index for elements of set
j	index for elements of set
k	index for elements of set
t	index for elements in set of sampling techniques
L	lower bound
U	upper bound

Parameters

K	size of initial sample set
K_{max}	maximum number of sample points
N	total number of input domain dimensions
M	total number of output domain dimensions

Continuous variables

x	vector of input/design variables
y	vector of output/response variables

Symbols

d_n	edge length of nth dimension of $\mathcal{D}$
d	vector of edge lengths of $\mathcal{D}$
$\mathcal{D}$	domain
Δ	departure function
ε	minimum allowed distance between two points
$\mathbb{E}$	expectation
f	computationally costly function
g_b	basis function in kriging
$\mathbb{N}$	set of natural numbers
ρ_k	kriging order
ρ_p	PRSM order
Q	test set size
$\mathcal{Q}$	test set
$\mathbb{R}$	set of real numbers
S	surrogate model form
$\mathcal{T}$	set of sampling techniques
$V_N(\mathcal{D})$	hyper-volume of $\mathcal{D}$
$\mathcal{X}_N^{(K)}$	N dimensional sample set of size K
$\mathcal{Y}_M^{(K)}$	M dimensional response set of size K
Z	random process

over the static ones viz. low computational expense and better approximation quality (Crombecq et al., 2011a). Typically, an adaptive sampling technique starts with a small set of sample points, and then adds points sequentially based on some user-defined criterion. Such criterion involves an objective (sometimes referred as a *score*) that aims to fill the domain (exploration) as well as improve the overall surrogate quality (exploitation) (Garud et al., 2017a; Crombecq et al., 2011a). We summarize various adaptive approaches from the literature and their vital characteristics like the exploration and exploitation criteria, dependence on the surrogate form, and the placement approach in Table 1. Although, we only discuss the key works from the adaptive sampling literature, Garud et al. (2017b) has dedicated an entire section for their discussion and the interested readers may refer to it for further details.

Jin et al. (2002) propose two approaches, namely the maximin scaled distance (MSD) and the cross validation (CV). The former is a modification of maximin distance based sampling that utilizes system information by assigning weights to the important variables while the latter uses CV error (Kohavi, 1995) to place new sample points. The CV approach can be viewed as a maximum sampling error approach with an additional feature of clustering constraint. Crombecq et al. (2009, 2011a) propose a novel and generic *score* based sequential strategy involving exploration and exploitation. They use a combination of derivative-based local linear approximations and Voronoi tessellations to place new sample points. Although the LOLA-Voronoi strategy has shown some promising results, it can be computationally intensive for large N. A recent work by Eason and Cremaschi (2014) proposes an adaptive sampling strategy for ANN surrogates. Instead of generating all sample points in one shot, they choose them gradually based on some score from randomly generated sample sets. The score considers the normalized nearest neighbor distance of a potential point from the current sample points and its normalized expected variance evaluated using jackknifing (Efron, 1982). Though their selection of sample points is systematic, it is still from randomly generated points. Cozad et al. (2014, 2015) propose an adaptive sampling for their surrogate modeling tool called ALAMO. They add sample points one at a time to the initial sample set. For each new sample point, they solve a derivative-free optimization problem to maximize the deviation of the surrogate from the real function. This can obviously be compute-intensive, as it requires the evaluation of the real function during optimization.

To this end, the adaptive sampling techniques in the literature can be broadly classified as either score-based or optimization-