A procedure for surrogate modelling using space-filling design of experiments and adaptive regression

Gints Jekabsons, Andis Lagzdins, Jurijs Lavendels
Institute of Applied Computer Systems, Riga Technical University
gintsj@cs.rtu.lv, andish@gmail.com, jurisl@cs.rtu.lv

ABSTRACT

In many industrial applications, to cut down either the cost of natural experiments or the computational cost of complex, high fidelity scientific and engineering simulations, they are often substituted with regression models (in this context also referred to as surrogate models) that mimic the behaviour of the original system as closely as possible while being much cheaper to evaluate. Primary objectives of surrogate modelling are to obtain a model that is as accurate as possible and to minimize the required computational and experimental effort, including minimizing the necessary number of sample points and utilizing an efficient modelling method. In this paper, a surrogate modelling procedure is proposed which incorporate: 1) a space-filling method for designing of experiments based on an analogy of uniform distribution of charged particles in alongside placed multidimensional phantom spaces; 2) an adaptive regression modelling method based on a heuristic search through an infinite model space.

INTRODUCTION

In many industrial applications, to cut down either the cost of natural experiments or the computational cost of complex, high fidelity scientific and engineering simulations, they are often substituted with regression models (in this context also referred to as surrogate models) that mimic the behaviour of the real world systems or the simulations as closely as possible while being much cheaper to evaluate (Chen et al., 2006; Kalnins et al., 2008; Simpson et al., 2001). Surrogate models are then used in place of further real world or simulated experiments making possible such routine tasks as design optimization, design space exploration, sensitivity analysis and what-if analysis which can require thousands or even millions experimental evaluations.

While building surrogate models, the exact, inner working of the system at hand is not assumed to be known (or even understood), solely the input-output behaviour is important. A model is constructed based on modelling the response of the system to a limited number of intelligently chosen sample points. The process of building a surrogate model usually involves two major steps which may be interleaved iteratively: 1) sample selection (known as design of experiments, DOE); 2) construction of the surrogate model that best describes the behaviour of the system and estimation of its predictive performance.

The primary objectives of surrogate modelling are to obtain a model that is as accurate as possible and to minimize the required computational and experimental effort (Jin et al., 2002; Kalnins et al., 2008). This includes minimizing the necessary number of sample points and utilizing a computationally efficient modelling method of high predictive performance.

In this paper, a surrogate modelling procedure is proposed which incorporate: 1) a space-filling method for designing of experiments based on an analogy of uniform distribution of charged particles in alongside placed multidimensional phantom spaces; 2) an adaptive regression model building method based on a heuristic search through an infinite space of models-candidates.

The rest of this paper is organized as follows. In the next section the space-filling method for designing of experiments is discussed. Then the approach of adaptive regression model building is described. And finally, the last section concludes the work and gives directions of possible future research.

DESIGN OF EXPERIMENTS

Important research issue associated with surrogate modelling is how to achieve good accuracy of a surrogate model with reasonable number of sample points. While the accuracy of a surrogate model is directly related to the modelling technique used and to properties of the problem itself, the type of sampling approaches have a direct influence on the approximation performance. It is generally accepted that space-filling designs, for example the Latin Hypercube design, are preferable for building of surrogate models. Currently, there is a wide range of literature concerning different methods for DOE including many approaches for space-filling designs (Jin et al., 2002; Chen et al., 2006; Auzins 2004; Santner, 2003).

One of the approaches for space-filling DOE is to use the analogy of abstract charged particles and their interaction forces. The approach is based on the assumption that charged particles enclosed in a finite space tend to arrive at a stable state, i.e., the system tends to take a state with minimal potential energy. As a result, a homogeneous experimental plan with uniformly deployed sample points throughout the space is obtained.

One of the first similar methods was proposed by Audze & Eglajs (1977). Their work was guided by the following simple considerations – it is known that gas molecules (without regard to its internal processes) in any enclosed space collocate steadily, supported by repulsion forces between molecules (see Figure 1). Forces between molecules make a field that has the potential energy and in any such system will eventually take a state in which potential energy is minimal. For simplicity, it is assumed that the repulsion forces between the molecules are inversely proportional to the square of the distance between them. It gives the following potential energy expression:

$$Q = \sum_{i=1}^{n-1} \sum_{j=i+1}^{1} \frac{1}{\sum_{k=1}^{d} (x_{ik} - x_{jk})^2} = \sum_{i=1}^{n-1} \sum_{j=i+1}^{1} \frac{1}{|x_i - x_j|^2}. \quad (1)$$

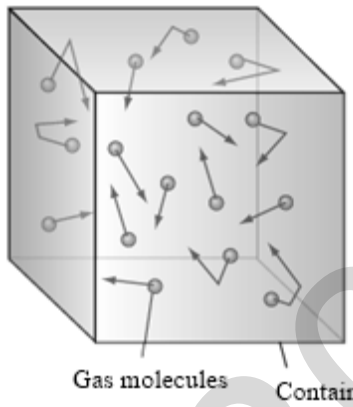

Fig. 1. Movement of gas molecules in a closed space

Molecules have their own internal processes, which make the molecules to move without interruption, therefore hereafter, instead of molecules, more abstract entities – charged particles, are discussed.. Charged particles interact with each other, depending on the distance at which they are located, as well as the magnitude of their charge.

Supplementing the above example with infinite space in all directions, where the actual domain of experiments with its experimental points is repeating over and over (hereinafter will be referred to as the phantom quasi-space and the phantom points therein), an infinite quasi-space with charged particles in it, is obtained. After a certain time, the system will take a stable position and the distance between all the charged particles will be constant. However, in practice to create an experimental design using this approach, the infinite set of quasi-spaces must be limited.

Using the idea of the phantom quasi-spaces ensures that the sample points have equal probability to stay in any position of area of the experimental domain. Without the phantom spaces a large subset of sample points would always end up exactly on the boundary of the experimental domain. This is especially important if the plan is highly dimensional or there are only very few sample points available. Figure 2 shows how the experiment plan is obtained without the phantom quasi-spaces, where P_i is ith sample point and F_i is resultant force.

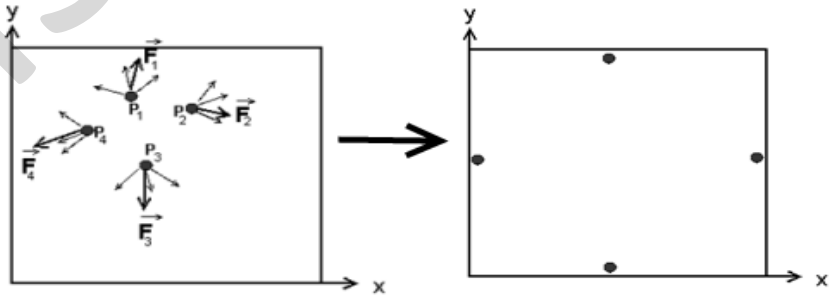

Fig. 2. Movement of charged particles without phantom spaces

A phantom space is identical to the area of actual domain of experiments with identical positions of the experiment points on it. Phantom spaces are distributed around the experimental area, creating a system that extends the domain, introducing with imaginary continuations. Because each phantom quasi-space also contains phantoms of each sample point, they interact with the actual sample points in the same way as the actual sample

points themselves. Consequently, it is ensured that the sample points are not concentrated close to the boundaries of the experimental domain, but instead are homogeneously distributed throughout the area.

In case of experimental design with one design parameter, there should be two phantom spaces – one before the actual experimental domain and one after the domain. Figure 3 shows how phantom spaces are spreading in the positive and negative directions. The phantom spaces can be imagined as infinite set of spaces in all possible directions (see Figure 4). Here, all experiment points and their phantoms are labelled as *x*.

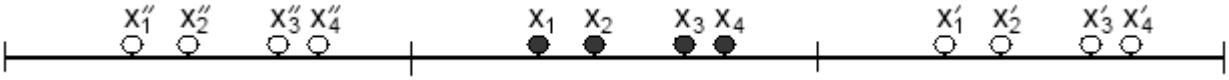

Fig. 3. Phantom spaces in case of one design parameter

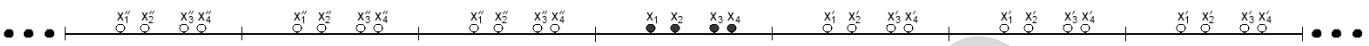

Fig. 4. Infinite phantom spaces in case of one design parameter

In two-dimensional case, it looks similarly. The infinite number of phantom quasi-spaces are spreading all around in all possible directions (see Figure 5). It can also be observed the rapidly growing number of the phantom spaces and phantom points depending on the design space dimensionality.

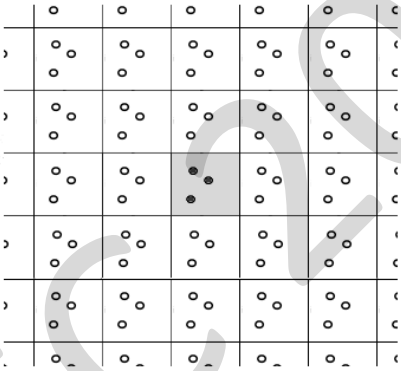

Fig. 5. Infinite set of phantom spaces in case of two design parameters

Each sample point has a certain number of phantoms depending on the dimensionality. Any actual sample point on each side of the single-level phantom spaces has symmetric phantom points with inversely directed charge vectors. As the forces acting between the charged particles are inversely proportion to the square of the distance between the particles, increasing the distance makes the difference between the opposing forces (symmetric) of distant-level phantom points become minimal:

$$\frac{1}{r^2} \approx \frac{1}{(r+\Delta)^2}, \; r >> \Delta, \tag{2}$$

where r is distance between a sample point and a distant phantom and Δ is the difference between distances from sample points to symmetric phantoms. Now the infinite set of phantom spaces can be reduced to a finite number of phantom quasi-spaces.

In order for the system to reach a stable state, the sample points (charged particles) must be allowed to move in all quasi-spaces. All the corresponding points in all the quasi-spaces move simultaneously identically to their counterparts and together with their corresponding actual sample points of the experimental domain. In a real world, all charged particles would move simultaneously and quickly obtain a steady state where the absolute values of resultant forces become minimal. However, in computer calculations, one must decide on a mechanism to determine how the charged particles must be relocated. In the practical implementation, in each step of the process a particle must be found that is affected from other points the most and move it in direction of resultant force. This way step-by-step the process continues until the system reaches a stable state. When the system is stabilized, it can be said that experiment plan has been created.

It should be noted that in this step-by-step process there are situations where the sample points from the actual domain area can move from the domain area to a phantom quasi-space. In this situation, as all the corresponding particles are moving simultaneously and identically, at the same time a phantom point is entering

the domain area from the opposite phantom quasi-space and becomes a sample point. And, as the number of the phantom quasi-spaces is infinite, there is no particle missing in any phantom quasi-space.

ADAPTIVE REGRESSION MODELLING

Originally surrogate modelling was associated with low-degree polynomial regression models which have global nature in describing numerical responses. They have been well accepted in engineering practice, as they require low number of sample points and are computationally very efficient. On other hand they are loosing efficiency when highly nonlinear behaviour should be approximated. Instead, higher-degree polynomials can be employed. However, if no special care is taken, they tend to overfit the data and produce high errors especially in regions where the sample points are relatively sparse.

One possible remedy for the overfitting problem is employment of the subset selection techniques. These are aimed to identify the best (or near best) subset of individual polynomial terms (basis functions) to include in the model while discarding the unnecessary ones, in this manner creating a sparse polynomial model of increased predictive performance.

However the approach of subset selection assumes that the chosen fixed full set of user-predefined basis functions (usually predefined just by fixing the maximal degree of a polynomial) contains a subset that is sufficient to describe the target relation sufficiently well. Hence the effectiveness of subset selection largely depends on whether or not the predefined set of basis functions contains such a subset. Generally, the required maximal degree is not known beforehand and needs to be guessed (or found by additional search over the whole subset selection process) since it will differ from one regression task to another. In many cases (especially when the studied data dependencies are complex and not well studied) this means either a non-trivial and long trial-and-error process or acceptance of a possibly inadequate model.

There exists a different approach for sparse polynomial model building – Adaptive Basis Function Construction, ABFC (Jekabsons, 2010; Jekabsons, 2008). The approach enables generating sparse polynomials of arbitrary complexity and degree without the requirement to predefine any basis functions or to pre-set the degree – all the required basis functions are constructed adaptively specifically for the data at hand. Additionally, in contrast to a number of other state-of-the-art surrogate modelling techniques the models built by the ABFC can be expressed as explicit and simple-to-use regression equations.

Assuming that x is an input to the actual computer analysis or natural experiment, generally a polynomial regression model can be defined as a basis function expansion:

$$F(x) = \sum_{i=1}^{k} \beta_i f_i(x) \tag{3}$$

where β is vector of coefficients for the model; k is the number of the basis functions included in the model; and $f_i(x)$ is a basis function generally defined as a product of original input variables each with an individual exponent:

$$f_i(x) = \prod_{j=1}^{d} x_j^{r_{ij}} \tag{4}$$

where d is the number of input variables; and r is a $k \times d$ matrix of non-negative integer exponents such that r_{ij} is the exponent of the jth variable in the ith basis function. Note that when for a particular basis function all the exponents are equal to zero, the basis function is the intercept term. The coefficients β are determined by minimizing least squares:

$$\beta = \arg\min_{\beta} \sum_{i=1}^{n} \left(F(x_{(i)}) - y_{(i)}\right)^2 \tag{5}$$

where n is the number of available sample points; $x_{(i)}$ is the input value of the ith sample point; and $y_{(i)}$ is the actual response value of the ith sample point.

Given a number of input variables d, matrix r with a specified number of rows k and with specified values of each of its elements completely defines the structure of a polynomial model with all its basis functions. Moreover, as neither the upper bounds of r elements' values nor the upper bound of k are defined, it is possible to generate polynomials of arbitrary complexity, i.e., of arbitrary number of basis functions each with arbitrary exponent for each input variable.

In order to efficiently build a sufficiently good regression model for a particular dataset, an efficient search mechanism is required enabling searching in an infinite space of polynomial models. In general, search mechanism of ABFC is organized as follows. The search is started from the simplest model – the model with one basis function which corresponds to the intercept term. New models are generated using so-called model refinement operators which enable adding, copying, modifying, and deleting the rows of *r*, i.e., adding, copying, modifying, and deleting the basis functions of the model (not only adding and deleting, as it is in subset selection methods).

The refinement operators can be categorized in two categories: "growers" and "purifiers". The growers do the main job – they "grow" the model. The purifiers on the other hand decrease the unnecessarily high exponents and delete the unnecessary basis functions. Without the use of simplification operators, a regression model may contain unnecessarily high exponents and include too many unnecessary basis functions, at the same time preventing truly necessary modifications (this is also known as the nesting effect (Pudil et al., 1994)) and increasing overfitting.

The initial state and the state transition operators together form a state space. Figure 6 shows a small example of a state space in ABFC when the number of input variables is three and all the four state transition operators are used. Each state represents a set of basis functions included in the regression model. The ordering of the states in the space is such that the simplest models and the simplest basis functions are reached first and, as the search goes on, increasingly complex models and basis functions can be reached.

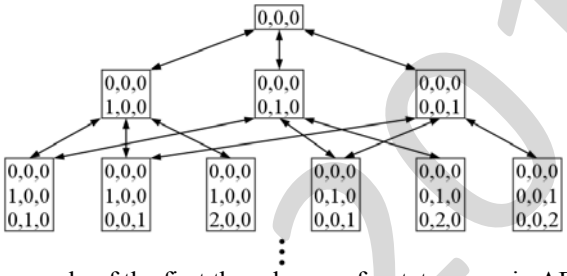

Fig. 6. A small example of the first three layers of a state space in ABFC when d = 3
(the space is infinite in the direction of more complex models)

Next, an efficient search strategy and a model evaluation measure are required. For ABFC, the search strategy of Sequential Floating Forward Selection (Pudil et. al., 1994) is adapted and the corrected Akaike's Information Criterion (AICC) (Hurvich & Tsai, 1989) is employed. The termination condition of the search process is met when the algorithm has generated a model which cannot be further refined using any of the refinement operators.

Additionally, in order to lower the general model building issues of selection bias and selection instability (Breiman, 1996; Jekabsons, 2008; Jekabsons, 2010; Loughrey, 2004, Cherkassky & Mulier, 2007), a technique of model averaging (also called ensembling or combining) is carried out. A typical model combination procedure consists of a two stage process (Cherkassky & Mulier, 2007). In the first stage, a number of different models are constructed. The parameters of these models are then held fixed. In the second stage, these individual models are linearly combined to produce the final model.

In ABFC, a Cross-Validation-like (CV) (Kohavi, 1995) resampling of the training data together with unweighted model averaging is employed. During resampling, the whole training data is randomly divided into *v* disjoint subsets (*v* typically being equal to 10). Then *v* overlapping training data sets are constructed by dropping out a different one of these *v* subsets. This produces *v* models built by *v* independent ABFC runs each using a different combination of CV partitioned data subsets. Next, the *v* models from the *v* CV iterations are combined using the unweighted model averaging. Note that, prior to combining, all the models are re-fitted to the whole training data set (without the CV partitioning). This is done to compensate for the smaller training sets used during the individual model building.

Model combining by unweighted model averaging consists in taking an unweighted average of predictions of all the models:

$$F_{comb} = \frac{1}{v}\sum_{i=1}^{v} F_i \ , \tag{6}$$

where F_i is *i*th individual model fro the *i*th CV iteration and F_{comb} is the combined model. For polynomial regression this simply means summation of all the polynomials and then a division of all the parameters of F_{comb} (that is also a polynomial) by *v*. Note that the parameter values of F_{comb} will not necessarily be optimal in the sense of the least squares loss (in fact they will be optimal only in special cases, e.g., when all F_i's are identical).

Figure 7 gives an outline of the ABFC model ensembling process when the number of CV folds v is three. Note however that for practical applications $v = 10$ is usually a better choice. This is because too small number of models in ensemble will yield too little diversity hindering the models to correct each others errors, but, on the other hand, using too many models will yield no further improvement.

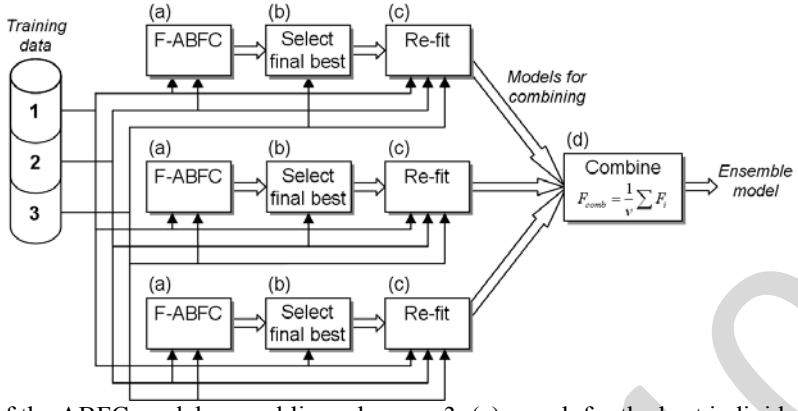

Fig. 7. An outline of the ABFC model ensembling when $v = 3$: (a) search for the best individual model; (b) select the one final best model; (c) re-fit the model (recalculate its parameters) using the whole training data; (d) combine the models

ABFC attempts to model arbitrary dependencies in data with little or no knowledge of the system under study. The user is normally not required to tune any hyperparameters. However, if there is sufficient additional domain knowledge outside the specific data at hand it may be appropriate to place some constraints on the final model. If the knowledge is fairly accurate, such constraints can improve the accuracy while saving computational resources.

For example the constraints might be one or more of the following: 1) limiting the maximal degree of all the basis functions (similar to the subset selection), i.e., $\forall i : 0 \le \sum_{j=1}^{d} r_{ij} \le p$; 2) limiting the maximal value of the exponent for each particular input variable in all the basis functions, i.e., $\forall i : 0 \le r_{ij} \le p_j$, where p_j is maximal exponent of the jth variable; 3) restricting contributions of specific input variables that are not likely to interact with others so that those variables can enter the model in basis functions only solely – with exponents of all other variables fixed to zero. These constraints, as well as far more sophisticated ones, can be easily incorporated in the ABFC.

A more complete discussion on the ABFC is given in (Jekabsons, 2010). A comparison of ABFC and other state-of-the-art surrogate modelling methods can bet found in (Jekabsons, 2010; Kalins et al., 2008; Kalnins, et al., 2009). ABFC implementation in Matlab is available at http://www.cs.rtu.lv/jekabsons/.

CONCLUSION

In this paper, a surrogate modelling procedure is proposed which incorporate: 1) a space-filling method for DOE based on an analogy of uniform distribution of charged particles in alongside placed multidimensional phantom spaces; 2) an adaptive regression model building method based on a heuristic search through an infinite space of models-candidates.

The proposed space-filling method is differs from most other similar methods in that it uses a notion of phantom spaces placed alongside the original experimental domain thereby equalizing the probabilities for the sample points being at any position in the domain.

The proposed adaptive regression modelling method offers adaptive heuristic search capabilities for building of surrogate models specifically for the data at hand. The approach is different from the standard subset selection approach in that it does not require user to guess the maximal degree of the models (or predefine the full set of basis functions). The modelling method itself constructs the basis functions necessary for creation of a model with adequate predictive performance.

Directions of future research include throughout empirical experiments for practical evaluation of efficiency of the proposed procedure.

REFERENCES

Audze, P., & Eglais, V. (1977). New approach for planning out of experiments. *Problems of Dynamics and Strength*, 35, Zinatne¸ Riga (in Russian), 104-107.

Auzins, J. (2004). Direct optimization of experimental designs. *10th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conf.*, AIAA paper, No. 2004-4578.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Annals of Statistics*, 24, 2350-2383.

Chen, V.C.P., Tsui, K-L., Barton, R.R. & Meckesheimer, M. (2006). A review on design, modeling and applications of computer experiments. *IIE Transactions*, 38(4), 273-291.

Cherkassky, V. & Mulier, F.M. (2007). *Learning from Data: Concepts, Theory, and Methods* (2nd ed.), Wiley-IEEE Press

Hurvich, C.M. & Tsai, C-L. (1989). Regression and Time Series Model Selection in Small Samples. *Biometrika*, 76, 297-307.

Jekabsons, G. (2008). Ensembling Adaptively Constructed Polynomial Regression Models. *International Journal of Intelligent Systems and Technologies (IJIST)*, 3(2), 56-61.

Jekabsons, G. (2010). Adaptive Basis Function Construction: an approach for adaptive building of sparse polynomial regression models. *Machine Learning*, Yagang Zhang (ed.), In-Tech, 127-156.

Jin, R, Chen, W., & Sudjianto, A. (2002). On sequential sampling for global metamodeling in engineering design. *Proceedings of DETC '02, ASME 2002 Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, DETC2002/DAC-34092, Montreal, Canada

Kalnins, K., Jekabsons, G. & Rikards, R. (2009, June). Metamodels for optimisation of post-buckling responses in full-scale composite structures. *Proceedings of 8th World Congress on Structural and Multidisciplinary Optimization*, Lisbon

Kalnins, K., Ozolins, O. & Jekabsons, G. (2008). Metamodels in design of GFRP composite stiffened deck structure. *Proceedings of 7th ASMO-UK/ISSMO international conference on engineering design optimization*, Association for Structural and Multidisciplinary Optimization in the UK, London, UK

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, San Mateo, CA, USA, 1137-1145.

Loughrey, J. & Cunningham, P. (2004). Overfitting in Wrapper-based Feature Subset Selection: the Harder You Try the Worse It Gets. *24th SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence*, 33-43.

Pudil, P., Ferri, F.J., Novovicova, J. & Kittler, J. (1994). Floating search methods for feature selection with nonmonotonic criterion functions. *Proceedings of the International Conference on Pattern Recognition*, 2, IEEE, Los Alamitos, CA, 279-283.

Santner, T.J., Williams, B.J., & Notz, W.I. (2003). The Design and Analysis of Computer Experiments, Springer

Simpson, T.W., Peplinski, J.D., Koch, P.N., & Allen, J.K. (2001). Metamodels for computer-based engineering design: survey and recommendations. *The Journal of Engineering with Computers*, Special Issue Honoring Professor Steven J. Fenves 17, 129–150.