Integration of Data Mining Into a Nonlinear Experimental Design Approach for Improved Performance

Guiying Zhang
Dept. of Food Science and Technology, University of California, One Shields Avenue, Davis, CA 95616
Dept. of Viticulture and Enology, University of California, One Shields Avenue, Davis, CA 95616

David E. Block
Dept. of Viticulture and Enology, University of California, One Shields Avenue, Davis, CA 95616
Dept. of Chemical Engineering and Materials Science, University of California,
One Shields Avenue, Davis, CA 95616

DOI 10.1002/aic.11955
Published online September 3, 2009 in Wiley InterScience (www.interscience.wiley.com).

Keywords: bioprocess engineering, data mining, decision tree analysis, optimization, nonlinear experimental design

Introduction

Recently, we have reported a nonlinear experimental design (*n*-DOE) method (Zhang et al., 2007)[1] that is more efficient than traditional optimization methods (e.g., explicit mathematical modeling techniques and statistical design of experiment (DOE) methods)[3–6] for the optimization of complex multidimensional systems (Zhang and Block, 2009).[2] The *n*-DOE method, RBFNN-TGA, includes a radial basis function neural network (NN) modeling technique and a novel search algorithm, a truncated genetic algorithm (TGA), to suggest a new set of experiments based on a developing knowledge base.[1] Having this optimization algorithm that can successfully deal with the dimensionality common in bioprocess optimization overcomes a potential limitation of most previously reported approaches to optimization.[7–9]

After evaluating this *n*-DOE method, we hypothesized that the algorithm could be further improved by including a data mining technique to ensure that only the critical variables of the dataset are used for the optimization, whereas unimportant variables that do not have significant effects on desirable responses/outputs are eliminated from the optimization process. We felt that this data mining addition would make the *n*-DOE more efficient for two reasons: (1) the NN model used in the algorithm would be more predictive with unimportant variables excluded[10,11] and (2) the size of the search space is reduced as the number of factors included in the optimization is reduced. Many data mining techniques capable of extracting critical features from a database have been reported.[10–14] In this work, we chose to process the variables using a decision tree analysis (DTA) technique since it is able to identify critical variables from a nonlinear database with both continuous and discrete variables,[10,11] before feeding the variables into a NN model. By integrating the DTA into our novel *n*-DO (termed n-DOE-DTA), our goal was to extend our original method[1] and decrease the number of experiments necessary for the *n*-DOE method to find an optimum, while maintaining its ability to find a true optimum. To evaluate this, we used a complex simulated optimization surface (a mathematical function) to assess the performance (i.e., effectiveness and efficiency) of the *n*-DOE-DTA and *n*-DOE approaches.

Methods

Response surface used

For optimization problems representative of the high nonlinearity and multidimensionality in practice, we created a 12-dimensional optimization surface

Correspondence concerning this article should be addressed to D. E. Block at deblock@ucdavis.edu

Current Address of Guiying Zhang: Amgen Inc., One Amgen Center Drive, Thousand Oaks, CA 91320.

Table 1. The Bounds and Type of Each Variable for a 12-Dimensional Optimization Surface

	Function (1)	
Variable (X)	Bounds	Parameter Type
x_1	[−2 2]	D
x_2	[−1 3]	C
x_3	[−5 5]	C
x_4	[−5 5]	D
x_5	[−1 1]	C
x_6	[−1 1]	D
x_7	[0 4]	C
x_8	[−2 2]	D
x_9	[−1 1]	C
x_{10}	[−1 1]	D
x_{11}	[−4 0]	C
x_{12}	[1 3]	D

C, continuous variable; D, discrete variable (only integer values used).

$$
\begin{aligned}
f(X) &= 100\left(x_2 - x_1^2\right)^2 + (1 - x_1)^2 \\
&+ \left(x_4^2 + x_3 - 11\right)^2 + \left(x_4 + x_3^2 - 7\right)^2 + x_4 + 3x_3 + 45 \\
&* 1.3356\left\{\begin{array}{l} 1.5(1 - x_5) + \exp(2x_5 - 1)\sin\left[3\pi(x_5 - 0.6)^2\right] \\ + \exp[3(x_6 - 0.5)]\sin\left[4\pi(x_6 - 0.9)^2\right] \end{array}\right\} \\
&+ 5\left(\log(x_7 + 1)^2\right)^2 + 3(x_8 - 1)^2 + 1.5x_9 \\
&+ \sin\left(3\pi(x_{10} - 0.6)^2\right) + (\exp(x_{11} - 1))^{0.8} - 0^{x_{12}} - 10.78
\end{aligned}
\qquad (1)
$$

to evaluate the performance of the new n-DOE-DTA optimization method and compare its performance with the n-DOE without integrated data mining. The bounds and types of variables for the function (1) are given in Table 1. Function (1) is composed of three complex two-dimensional functions (modified Rosenbrock function, modified Himmelblau function, and Maechler Additive function)[5] and six other nonlinear terms, in which x_{12} is a random variable, the average output over the constrained input space is 657, the global minimum is 0 at the coordinates (1, 1, −3.32, −4, 0.655, 0, 0, 1, −1, −1, −4, random), and the global maximum is 3701.5 at the point of (−2, −1, 5, 5, −1, 1, 4, −2, 1, 1, 0, random). The 12-dimensional function was chosen so that x_1 is the most critical parameter in determining the response of the function with x_2 next important and so forth down to x_{12}. Five-percent simulated noise was added to the responses of the optimization surfaces.[1]

Decision tree analysis

Using different sizes of simulated datasets from function (1), we have examined several widely used decision tree algorithms based on different split criteria, including ID3 (based on information gain),[15] C5 (a descendant of C4.5, based on information gain ratio),[16] CART (based on Gini index, GI),[17] and P_0 null hypothesis probability (based on a probability metric).[18] By comparing the algorithms, we found that a GI-based DTA can generally predict more significant attributes from a small dataset (data not shown). Therefore, in this work, the GI split criterion was then used to select the split attribute to construct decision trees. Theoretical details on the GI can be found in earlier publications.[17,19]

n-DOE method integrated with a DTA technique (n-DOE-DTA)

A fractional factorial design method, a minimum run equireplicated resolution IV design,[20,21] was used to generate initial datasets (namely, the first batch of experiments) for the n-DOE-DTA and n-DOE approaches.

The n-DOE method (Figure 1a) is discussed in detail in our previous report.[1] A flow chart of the n-DOE method integrated with a DTA technique (i.e., n-DOE-DTA) that we devised in this work is shown in Figure 1b and can be contrasted with the n-DOE approach. For the n-DOE-DTA, a GI-based DTA technique is applied to a database to identify the important features (variables). The extracted critical variables are then fed into a NN for a better model, which will eventually lead to a better optimum by iterating the n-DOE method. We used three previously reported success indicators to evaluate the performance of the n-DOE and n-DOE-DTA approaches.[1] These indicators are the mean output of optimum conditions identified by the algorithm (mean$_{\text{opt}}$), the mean number of experiments needed to locate an optimum, and the successful decrease percentage in terms of mean$_{\text{opt}}$ (SDP$_{\text{mean}}$), which is a measure of how close the algorithm comes to finding the true optimum.

The computing environment and statistical software for this work can be referred to Zhang and Block (2009).[2] (Zhang and Block, submitted).

Results

To examine this integrated approach, we used the complex 12-dimensional function in function (1) to assess the performance of the n-DOE-DTA method and compared it with that of the n-DOE method. The optimization goal is the minimum of the function. A dataset with 5% noise over responses[1] was simulated using this function in the form of a minimum run equireplicated resolution IV design.[20,21] This dataset containing 26 experimental points [with a minimum of 543.3 at the input coordinates of (2, 3, −5, −5, 1, 1, 0, −2, −1, 1, 0, random)] was used as the initial dataset of the n-DOE-DTA and n-DOE approaches. For each new batch of experiments, eight new experimental points were suggested for the next batch of experimentation for each of the two approaches. Both algorithms were run five times to evaluate algorithm variability.

For the n-DOE-DTA method, a GI-based DTA technique was applied to the 12-dimensional initial simulated dataset in the first batch of optimization to assess the ability of this technique to choose the most important variables. Figure 2 is the decision tree generated from the 12-dimensional initial dataset using the DTA technique. As shown in the tree, the first split chosen at the root is x_2 (a critical variable), but not the most critical variable of this function, x_1. This tree then identified two critical variables (x_3, x_4) in the second level, a critical variable (x_6) in the third level, along with the most critical variable (x_1), a critical variable (x_5), and a less critical variable (x_7) in the fourth level. In implementing the n-DOE-DTA, we were able to eliminate five less significant variables (x_8, x_9, x_{10}, x_{11}, x_{12}) from the optimization using

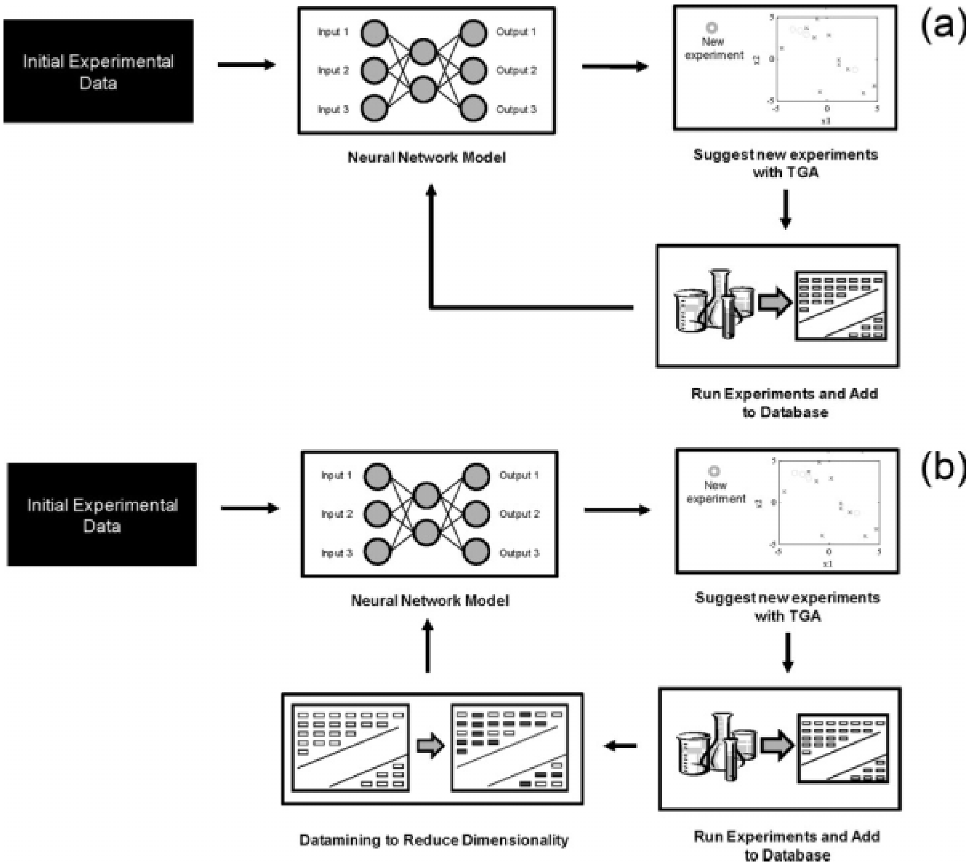
(a)
Initial Experimental Data
Input 1
Input 2
Input 3
Output 1
Output 2
Output 3
Neural Network Model
New experiment
x2
x1
Suggest new experiments with TGA
Run Experiments and Add to Database
(b)
Initial Experimental Data
Input 1
Input 2
Input 3
Output 1
Output 2
Output 3
Neural Network Model
New experiment
x2
x1
Suggest new experiments with TGA
Run Experiments and Add to Database
Datamining to Reduce Dimensionality

Figure 1. Comparison of the two n-DOE methods assessed.

(a) An illustration of the previous *n*-DOE developed and (b) the new method in which critical inputs are extracted using a database mining technique (e.g., DTA) and fed into neural network for a better NN model. The latter steps allow us to further reduce the necessary experiments and achieve a better-targeted output.

just the initial data set as these variables did not appear anywhere in the DTA of the initial dataset.

The cycle of NN model building and TGA selection is repeated, with the NN model accumulating knowledge in each cycle just as an experienced researcher would. A corresponding new DTA was constructed after each subsequent round of eight experiments to find if more factors could be eliminated dynamically during the experimentation. Even after

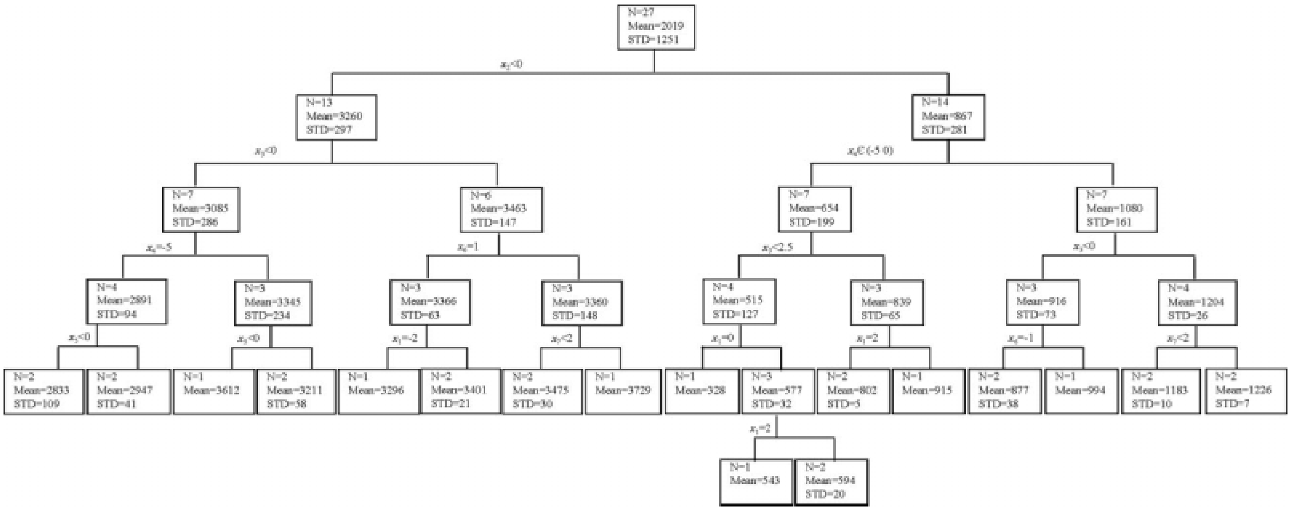

Figure 2. Decision tree generated from the initial dataset of the 12-dim function (1).

The initial dataset was generated by the minimum run equireplicated resolution IV design. Five variables (x_8, x_9, x_{10}, x_{11}, x_{12}) were removed from the dataset at batch 1, whereas x_7 was further removed from the tree at batch 5. *N* is the number of experiments, and STD is the standard deviation.

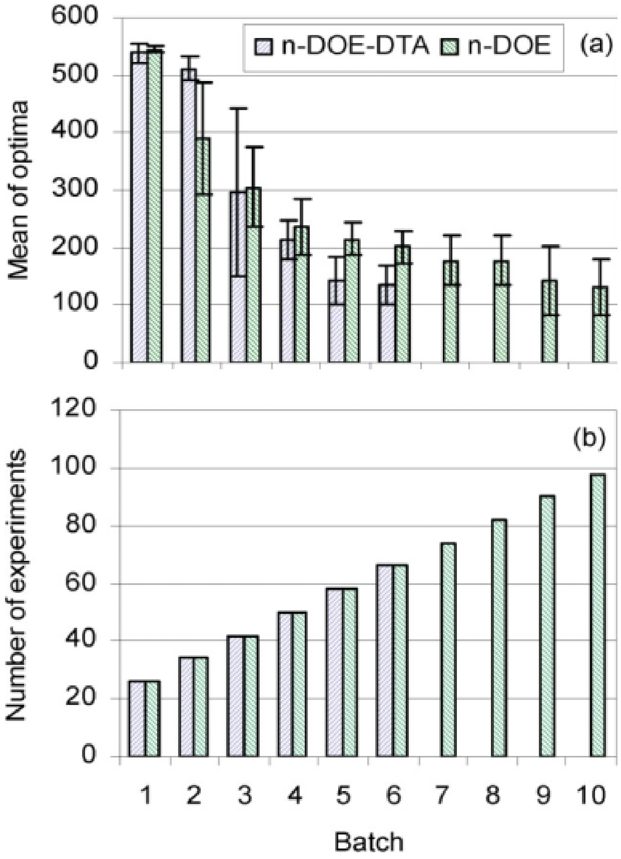

Figure 3. Comparison of the performance of the *n*-DOE-DTA and *n*-DOE approaches using an optimization surface, Function (1).

The iteration of the *n*-DOE-DTA method was ceased earlier (i.e., at batch 6) since no significant further improvement in the minimum was found in two consecutive batches, i.e., batch 5 and batch 6. This illustrates that the *n*-DOE-DTA method reaches the optimum in 32 fewer experiments. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

elimination of factors, the responses of new experiments were still generated using the 12-dimensional function (1), with values of the critical variables suggested by the DTA technique and the values of the eliminated insignificant variables chosen to be the optimum from the initial dataset. Using this iterative approach, factor x_7 was further eliminated at batch 5 as it was no longer identified as important by the DTA.

Figure 3 shows the optimization results of the *n*-DOE-DTA method for function (1). We ceased the iterations of this algorithm at batch 6, since no significant improvement over the optimum was achieved (i.e., a mean_{opt} at batch 5 of 142 ± 41 compared with a mean_{opt} at batch 6 of 134 ± 33). The optimization results for the control *n*-DOE method are also shown in Figure 3. We terminated the simulation at batch 10, since the mean_{opt} (132 ± 50) of the *n*-DOE method at batch 10 is very close to that of the *n*-DOE-DTA method at batch 6.

As can be seen in Figure 3, the optimum (minimum) of each batch generally decreased as the batches/experiments increased. Comparing the results of the two approaches for this complex function, the *n*-DOE-DTA method found a minimum of 134 ± 33 and a SDP_{mean} of 79.6% in 66 experiments, whereas the *n*-DOE method required 98 experiments to achieve a similar minimum of 132 ± 50 and a SDP_{mean} of 79.9%.

Conclusion

We have already shown that the *n*-DOE without data mining represents a significant improvement over common traditional statistical approaches to experimental optimization.[1] Here, we have shown that the incorporation of a DTA technique into our novel *n*-DOE method can dynamically reduce the dimensions of a complex dataset and remove insignificant variables leading to even more efficient experimental optimization of real complex processes.

Acknowledgements

This work was supported by the California Dairy Research Foundation and the University of California Discovery Grant program.

Notation

DOE	= design of experiment
DTA	= decision tree analysis
GI	= Gini index
$mean_{opt}$	= the mean of all optima located from NN models
n-DOE	= nonlinear experimental design
n-DOE-DTA	= an *n*-DOE method integrated with a DTA technique
NN	= neural network
RBFNN-TGA	= radial basis function neural network-truncated genetic algorithm based experimental design method
TGA	= truncated genetic algorithm
SDP_{mean}	= the SDP in terms of $mean_{opt}$

Literature Cited

1. Zhang G, Olsen MM, Block DE. New experimental design method for highly nonlinear and dimensional processes. *AIChE*. 2007;53: 2013–2025.
2. Zhang G, Block DE. Using highly-efficient nonlinear experimental design methods for optimization of Lactococcus lactis fermentation in chemically-defined media. *Biotechnology Progress*. 2009; (in press).
3. Zhang G, Mills DA, Block DE. Development of chemically-defined media supporting high cell density growth of lactococci, enterococci, and streptococci. *Appl Environ Microbiol*. 2009;75:1080–1087.
4. Schepers AW, Thibault J, Lacroix C. *Lactobacillus helveticus* growth and lactic acid production during pH-controlled batch cultures in whey permeate/yeast extract medium. I. Multiple factor kinetic analysis. *Enzyme Microbial Technol*. 2002;30:176–186.
5. Schepers AW, Thibault J, Lacroix C. *Lactobacillus helveticus* growth and lactic acid production during pH-controlled batch cultures in whey permeate/yeast extract medium. II. Kinetic modeling and model validation. *Enzyme Microbial Technol*. 2002;30:187–194.
6. Li C, Bai JH, Cai ZL, Fan OY. Optimization of a cultural medium for bacteriocin production by *Lactococcus lactis* using response surface methodology. *J Biotechnol*. 2002;93:27–34.
7. Weusterbotz D, Pramatarova V, Spassov G, Wandrey C. Use of a genetic algorithm in the development of a synthetic growth-medium for Arthrobacter Simplex with high hydrocortisone delta(1)-dehydrogenase activity. *J Chem Technol Biotechnol*. 1995;64:386–392.
8. Franco-Lara E, Link H, Weuster-Botz D. Evaluation of artificial neural networks for modelling and optimization of medium composition with a genetic algorithm. *Process Biochem*. 2006;41:2200–2206.
9. Bapat PM, Wangikar PP. Optimization of Rifamycin B fermentation in shake flasks via a machine-learning-based approach. *Biotechnol Bioeng*. 2004;86:201–208.
10. Buck KKS, Subramanian V, Block DE. Identification of critical batch operating parameters in fed-batch recombinant *E. coli*

Fermentations using decision tree analysis. *Biotechnol Prog*. 2002; 18:1366–1376.

11. Subramanian V, Buck KKS, Block DE. Use of decision tree analysis for determination of critical enological and viticultural processing parameters in historical databases. *Am J Enol Vitic*. 2001;52:175–184.
12. Coleman MC, Buck KKS, Block DE. An integrated approach to optimization of *Escherichia coli* fermentations using historical data. *Biotechnol Bioeng*. 2003;84:274–285.
13. Samuels ML, Witmer JA. *Statistics for the Life Sciences*, 3rd ed. New Jersey: Pearson Education, Inc., 2003.
14. Duda RO, Hart PE, Stork DG. *Pattern Classification*, 2nd ed. New York: Wiley, 2001.
15. Quinlan JR. Induction of decision trees. *Mach Learn*. 1986;1:81–106.
16. Quinlan JR. *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann, 1993.
17. Breiman L, Friedman J, Olshen R, Stone C. *Classification and Regression Trees*. Belmont, CA: Wadsworth International Group, 1984.
18. Martin JK. An exact probability metric for decision tree splitting and stopping. *Mach Learn*. 1997;28:257–291.
19. Raileanu LE, Stoffel K. Theoretical comparison between the Gini Index and Information Gain criteria. *Ann Math Artif Intell*. 2004;41: 77–93.
20. *Design-Expert7.1 User's Guide: Two-Level Factorial Tutorial*. Minneapolis, MN: Stat-Easy, Inc., 2007.
21. Box GEP, Hunter WG, Hunter JS. *Statistics for Experimenters*. New York: Wiley, 1978.

Manuscript received June 13, 2008, and revision received Feb. 24, 2009.