Integrating data mining to robust design procedure for a large data set and noise factors

Vo Thanh Nha, Nguyen Khoa Viet Truong, Sangmun Shin*
Department of Systems Management & Engineering, Inje University, Gimhae, KN 621-749, South Korea

Abstract

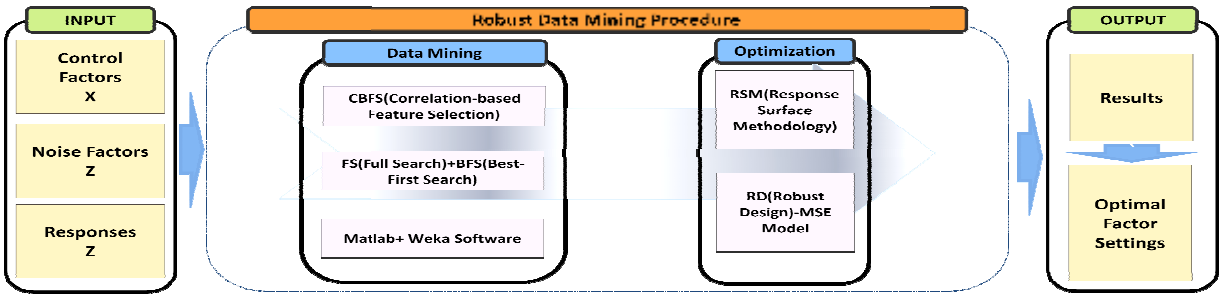
Although robust design (RD) data mining (DM) have received consistent attention as separate research fields from science and engineering researchers and their associated community for more than twenty years, there is room for improvement. In many industrial situations associated with data analysis, a large data set often includes a large number of input control factors as well as noise factors. In order to address this issue, the primary objective of this paper is to propose a new integrated DM-RD procedure for a large data set including a large number of input control and noise factors. The proposed method focuses on four main issues. First, in order to address dimensionality problems regard as a large number of input factors, we utilized a data mining (DM) method called correlation based feature selection (CBFS) to find significant factors. Second, for effective search methods, we consider best first search (BFS) and full search (FS) algorithms by conducting comparative analyses of these two algorithms. Third, based on the dimensionality reduction (i.e., significant factor selections) from the DM step, we then perform further statistical analyses, such as response surface methodology (RSM) and RD to obtain the optimal factor settings and statistical inferences for both control and noise factors. Finally, a numerical example and a comparative study are conducted for verification purposes.

Keywords:
Robust Design, Data Mining, Correlation-Based Feature Selection, Response Surface Methodology

1. Introduction

Continuous development and integrating data mining to robust design have become widely recognized by the industry researchers. In 1999, Hall M.A developed an method to find the merit of the subset [10]. Shin et. al [5-7,11] developed an integrated procedure combining a DM method and Taguchi methods. The factor selection algorithm performs a search through the space of feature subsets [3]. In general, two categories of the algorithm have been proposed to solve the factor selection problem. The first category is based on a filter approach that is independent of learning algorithms and serves as a filter to sieve the irrelevant factors. The second category is based on a wrapper approach, which uses an induction algorithm itself as part of the function evaluating factor subset [4]. Existing studies in DM mostly focus on finding patterns in large data sets and further using them for organizational decision making. DM methods also may not discuss the robustness of solutions, either by considering data preprocesses for outliers, or by considering uncontrollable noise factors. In order to address this issue, the primary objective of this paper is to propose a new integrated DM-RD procedure for a large data set including a large number of input control and noise factors. The proposed method focuses on four main issues. First, in order to address dimensionality problems regard as a large number of input factors, we utilized a data mining (DM) method called correlation based feature selection (CBFS) to find significant factors. In this step, we used matlab and weka software for calculate correlations between factor to factor and factor to response. Second, for effective search methods, we consider best first search (BFS) and full search (FS) algorithms by conducting comparative analyses of these two algorithms. By using matlab and weka software we can find the best subset based on correlation between factor and factor, factor and response. Third, based on the dimensionality reduction (i.e., significant factor selections) from the DM step, we then perform further statistical analyses, such as response surface methodology (RSM) and RD to obtain the optimal factor settings and statistical inferences for both control and noise factors. In optimization model, we use mean

square error (MSE) to find optimal setting. Finally, a numerical example and a comparative study are conducted for verification purposes. An overview of the proposed integrated DM-RD procedure can be illustrated in Fig. 1.



Figure[1]. An overview of the integrated DM-RD procedure

2. Proposed method

2.1. Pre-Processing data

2.1.1. Correlation-Based Feature Selection(CBFS) method

Correlation-based feature selection (CBFS) is a filter algorithm that ranks subsets of input features according to a correlation-based heuristic evaluation function. The bias of the evaluation function is toward subsets that contain a number of input factors, which are not only highly correlated with a specified response but also uncorrelated with each other ((Langley (1994); Xu and Kammel (2004); Shin (2007)). Among input factors, irrelevant factors should be ignored because they may have low correlation with a given a response, even though some selected factors are highly correlated with one or more of these selected factors. The acceptance of a factor depends on the extent to which it predicts the response in areas of interest that have not already been predicted by other factors. The evaluation function of the proposed subset is

$$= \quad / \quad + \quad (\quad -1) \qquad (1)$$

Where , and represent the heuristic evaluation value of a factor subset containing factors, the mean of factor-response correlation($\in$), and the mean of factor-factor inter-correlation, respectively. $+ \quad (\quad -1)$ and indicate the prediction of the response based on a set of factors and the redundancy among the factors.

In order to measure the correlation between two factors, or a factors and the response, we use correlation between two random variable and . If we have a series of n measurements of and written as and where $= 1,2, \ldots,$, then the sample correlation coefficient, can be used to estimate the population Pearson correlation r between and . The sample correlation coefficient is written $= \sum \quad (\quad - \quad) (\quad - \quad) \quad (\quad -1)$, where $\bar{}$ and are the sample mean of and , and are the sample standard deviation of and .

2.1.2. Full Search Algorithm

In order to finding a best subset , one of the most effective methods is the full search method, which is any algorithm that follows the problem solving metaheuristic of making the locally optimal choice at each stage with the hope of finding the global optimum. In cases, with factors less than one-hundred, we can enumerated all subsets and base on the evaluation function given in Equation (1) is a fundamental element of CBFS for imposing a specific ranking on factor subsets in the search spaces. We show the process of finding best subset algorithm by the following steps:

Table[1] The procedure of greedy search algorithm (GSA)

Step 1. Begin with the OPEN list containing the start state, the CLOSE list empty, and $\leftarrow 0$, $\leftarrow$ (put

start state to).
Step 2. For each next subset of ,
Step 3. If ( ) ≥ , then ← ( ), ←
Step 4. Return ,

2.2. Robust Design

2.2.1. Response surface methodology (RSM) based on noise factors

RSM is a statistical tool that is useful for modeling and analysis in situations where the response of interest is affected by several factors. RSM is typically used to optimize the response by estimating an input-response functional form when the exact functional relationship is not known or is very complicated. For a comprehensive presentation of RSM, Box et al.[5] and Shin and Cho[12] provide insightful comments on the current status and future direction of RSM. In many industrial situations, a manufacturing or service process contains both control and noise factors which may not be addressed.[9] Supposing that there are k controllable factors $x = [x_1, x_2, \ldots, x_k]$, and noise factors $z = [z_1, z_2, \ldots, z_k]$, the response model incorporating both control and noise factors is given by

$$(,) = + + + +$$

$$= ( ) + h(,) + \psi \quad (2)$$

Where = ⋮ , z = ⋮ , = ⋮ , b= ⋮

$$= \begin{matrix} & /2 & \cdots & /2 \\ /2 & & \cdots & /2 \\ \vdots & \vdots & \ddots & \vdots \\ /2 & /2 & \cdots & \end{matrix} \qquad = \begin{bmatrix} & /2 & \cdots & /2 \\ /2 & & \cdots & /2 \\ \vdots & \vdots & \ddots & \vdots \\ /2 & /2 & \cdots & \end{bmatrix}$$

where , , , and denote the vector forms of estimated regression coefficients associated with the main effects of control and noise factors, and the matrix forms of estimated regression coefficients associated with quadratic and interaction effects of control and noise factors, respectively. f(x), h(x, z), and ψ are the portion of the model that involves only the control factors, the term involving the main effects of the noise factors and the interactions between the control and noise factors, and a random error which is assumed to have a normal distribution with zero mean and certain variance, respectively. The detailed calculation of h(x, z) is

$$h(,) = + \quad (3)$$

By taking the expectation of the response model in Eq. (2), the mean response model can be derived as follows:

$$[(,)] = ( ) = + + \quad (4)$$

$$[(,)] = [+] + \quad (5)$$

Denoting variance of the noise factor as and assuming that the noise factor and the random errors ε have zero covariance, is the mean square error on analysis of variance (ANOVA).

2.2.1. Robust Design Model

RD methods, based on the concept of building quality into a design, are increasing popular in industry, primarily because of their simplicity and practicality. The method has been extensively applied in process design and is typically used to determine close to optimal operating conditions for process parameters. RD has received much attention from researchers and practitioners, and there has been an intense research effort to improve its applicability. RD was acknowledged as Taguchi's most significant contribution, for his two-step procedure that minimizes variability subject to a zero bias constraint. In this paper, we investigate several design optimization issues to further facilitate the RD principles.

By modeling process mean and variance separately, based on experimental data, the dual response approach achieves the goal of RD by minimizing the mean and the variance at the target as follows

$$(\quad [\ (\ , \ )] - \ ) + \quad [\ (\ , \ )] \qquad (6)$$

$$\in \Omega$$

where $\ , \Omega$ denote the target value for the mean response, the domain of control factors, respectively.

3. Numerical example

To effectively demonstrate the implementation of our proposed methodology, a real case study of a process to produce a placebo tablet has been conducted in which a number of design variables were considered. The data used in this numerical example were obtained from a continuous real-time tablet manufacturing process. The tablet manufacturing process is classified into three stages, namely: flow, compression, and ejection process. The objective of this study is to optimize commonly each desired bias and variability of tablet quality characteristics including hardness (y). Then, based on prior information about the system under investigation, it logically follows that first pressure (x_1) to remove air in the granules, second pressure (x_2) to produce tablets, first dwell time (x_3) to remove air in the granules, second dwell time (x_4) to produce tablets, speed to remove first a punch (x_5), speed to remove second a punch (x_6), speed to eject tablets (x_7), the amount of overfill (x_8), the amount of dust (x_9), and particle size (x_{10}) are the control factors and humidity (z_1) and temperature (z_2) are the noise factors considered in this study. In this particular case, the quality characteristics of interest have conflicting objectives as shown in Table 2.

Table[2]. Quality characteristics of hardness (H)

runs													
1	4.575	29.694	0.765	2.741	0.269	0.294	0.343	13.230	0.134	134.460	52.080	20.820	38.788
2	4.485	38.514	0.750	3.819	0.264	0.474	0.477	21.330	0.118	139.320	41.280	27.420	67.237
…	…	…	…	…	…	…	…	…	…	…	…	…	…
99	5.113	32.928	0.855	5.328	0.301	0.672	0.666	30.240	0.143	158.760	59.520	25.020	73.103
100	5.023	32.781	0.840	3.024	0.295	0.318	0.378	14.310	0.147	148.230	57.360	22.980	47.030

As shown in Table 3 and Table 4, DM results demonstrate three, four significant factors among twelve factors using Full Search and Best First Search, respectively. Base on the result we can see the merit of best subsets found using Full Search better than the merit of best subset using Best First Seach. Also we have three factors by using Full Search and four factors by using Best First Search.

Table [3] Illustrates DM results using Full Search Algorithm (MATLAB)

Select Evaluator	The response attribute Selected attributes	, ,
Search Method	Full Search	Full Search forward
	Total number of subsets evaluated	2^{12} - 1
Merit Of best subsets found	0.7992	

Table [4] Illustrates DM results using Best First Search Algorithm (WEKA software)

Select Evaluator	The response attribute Selected attributes	 , , ,
Search Method	Best First Search	Best First forward
	Total number of subsets evaluated	79
Merit Of best subsets found.	0.767	

Based on the results of the significant factor selection, RSM was performed using the MINITAB to identify comprehensive relationships among a large number of factors and their associated responses. As shown in Table 5,6 the RSM result using MINITAB. Furthermore , the P-Value of factors significant including , , .

Table [5] RSM result using Full Search Algorithm

Predictors	Coef	SE Coef	T	P
Constant	178.00	209.63	0.85	0.398
x2	006.30	4.26	1.47	0.144
x9	-4397.70	2169.29	-2.03	0.046
z2	-0.00	9.02	-0.01	0.998
x2*x2	0.10	0.06	2.43	0.017
x9*x9	15340.90	6465.51	2.37	0.020
x2*x9	-48.70	36.27	-1.34	0.183
x2*z2	-0.30	0.12	-2.76	0.007
x9*z2	99.70	54.13	1.84	0.069
S = 8.016	R-Sq = 73.1%	R-Sq(adj) = 70.7%		

Analysis of Variance

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Regression	8	15864.60	15864.56	1983.07	30.86	0.000
Linear	3	14442.70	756.67	252.23	3.92	0.011
Square	2	400.90	796.83	398.42	6.20	0.003
Interaction	3	1020.90	1020.92	340.31	5.30	0.002
Residual Error	91	5847.90	5847.92	64.26		
Total	99	21712.50				

Table [6] RSM result using Best First Search Algorithm

Predictors	Coef	SE Coef	T	P
Constant	-152.80	262.12	-0.583	0.561
x1	132.80	62.67	2.118	0.037
x2	-1.40	6.85	-0.197	0.844
x9	-1653.30	2294.22	-0.721	0.473
z2	-3.20	10.09	-0.319	0.750
x1*x1	-6.80	6.44	-1.056	0.294
x2*x2	0.20	0.06	2.732	0.008
x9*x9	25827.40	6982.11	3.699	0.000
x1*x2	3.20	1.69	1.902	0.061
x1*x9	-1269.00	390.48	-3.250	0.002
x1*z2	-0.00	1.83	-0.023	0.982
x2*x9	-90.80	40.01	-2.268	0.026
x2*z2	-0.50	0.14	-3.406	0.001
x9*z2	166.90	56.50	2.954	0.004
S = 7.562	R-Sq = 77.3%	R-Sq(adj) = 73.9%		

Analysis of Variance						
Source	DF	Seq SS	Adj SS	Adj MS	F	P
Regression	13	16794.4	16794.37	1291.875	22.59	0.000
Linear	4	14463.6	359.23	89.808	1.57	0.189
Square	3	589.7	1284.72	428.240	7.49	0.000
Interaction	6	1741.1	1741.11	290.184	5.07	0.000
Residual Error	86	4918.1	4918.11	57.187		
Total	99	21712.5				

After that, Using Eq.(2), the estimated response function can be obtained by

$$(,) = \quad + \quad + \quad + \quad +$$

where $= \quad = \quad , \quad = [\ ] , \quad = \begin{matrix} 6.3 \\ -4397.7 \end{matrix} ,$ b= $[-0.023]$

$$= \begin{matrix} 0.137 & -24.357 \\ -24.357 & 15340.9 \end{matrix} \qquad = \begin{matrix} -0.3286 \\ 99.717 \end{matrix}$$

And using Eqs. (4) and (5), the response function of the mean and variance can then be calculated by

$$[(,)] = ( ) = \quad + \quad +$$

$$[(,)] = \quad [+ \quad] +$$

where $= 6.335$, $= 59.673$ using the result of ANOVA. The estimated process mean and the variance functions are applied to the RD optimization models, including MSE model. In order to reduce the process variation, the target value of is given that $\tau = 50$. Table 7 provides RD optimization results using MSE models. Base on the result we can see best first search better than full search algorithm.

Table [7] RD optimization results using MSE model

RDM Approaches	Optimization factors Settings (*)			Optimization parameters		MSE
	*	*	*	^(*)	(*)	
BFS	4.112	30.474	0.110	50.000	57.187	57.187
FSA	N/A	29.283	0.100	50.072	60.196	64.791

4. Conclusions

In this paper, we focused on four fold. First, in order to address dimensionality problems regard as a large number of input factors, we utilized a data mining (DM) method called correlation based feature selection (CBFS) to find significant factors. Second, for effective search methods, we consider best first search (BFS) and full search (FS) algorithms by conducting comparative analyses of these two algorithms. Third, based on the dimensionality reduction (i.e., significant factor selections) from the DM step, we then perform further statistical analyses, such as response surface methodology (RSM) and RD to obtain the optimal factor settings and statistical inferences for both control and noise factors. We finally showed that full search is better than best first search algorithm when we find the best subset. But optimization result using best first

search is better than full search. The consideration of different way to find best subset can be a possible future research issue.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (20100104)

References

1. P. Tan, M. Steinbach, and V. Kumar., 2006, Introduction to data mining, Pearson Education Inc. Boston, USA.
2. C.T. Su, M.C. Chen, and H.L. Chan., 2005, "Applying Neural Network and Scatter Search to Optimize Parameter Design with Dynamic Characteristics" Journal of the Operational Research Society, Vol.56, pp.1132-1140.
3. D. Allen., 1974, "The Relationship between Variable Selection and Data Augmentation and a Method for Prediction," Technometrics, Vol.16, pp.125–127.
4. P. Langley., 1994, "Selection of Relevant Features in Machine Learning" Proceedings of the AAAI Fall Symposium on Relevance, pp. 140-144.
5. S. Shin and B.R. Cho., 2005, "Bias-specified robust design optimization and its analytical solutions," Computer & Industrial Engineering, Vol.48, pp.129-140.
6. S. Shin, G. Yi, Y. S. Choi., 2007, M. G. Choi and C. Kim, Development of a robust data mining method using CBFS and RSM, Lecture Notes in Computer Science 4378 (2007) 377–388
7. Yang. L, S.Shin, Choi.Y, Park. K, Kaewkuekool.S, Chantrasa, R. & Lila. B, (2007). Development of an Extended Robust Data Mining (ERDM) Model, Proceedings of International Conference on Control, Automation and Systems, pp. 1523-1528, ISBN:978-89-950038-6-2, Seoul, Korea, Oct. 2007, IEEE Press, New York.
8. G. E. P. Box, S. Bisgaard and C. Fung, An explanation and critique of Taguchi's contributions to quality engineering, International Journal of Reliability Management 4 (1988) 123–131.
9. D. C. Montgomery, Introduction to Statistical Quality Control 4th edition. (John Wiley& Sons, New York, 2001).
10. Hall, M. A., 1999: Correlation-based Feature Selection for Machine Learning. Ph.D diss.Waikato University, Department of Computer Science. Hamilton, New Zealand.
11. S.Shin, K.Park, Byung-Nam Kim., 2008, "Development of Robust Data Computing Methodology (RDCM) for a Multidisciplinary Pharmaceutical Process Design," icicic, pp.256, 2008 3rd International Conference on Innovative Computing Information and Control.
12. H. Jeong, S. Shin*, and B.R. Cho., (2008), "Integrating Data Mining to a Process Design Using the Robust Bayesian Approach," International Journal of Quality, Reliability and Safety Engineering, 15 (5), 441-464.