Available online at www.sciencedirect.com

ScienceDirect

Procedia Manufacturing 39 (2019) 618–624

Procedia MANUFACTURING

www.elsevier.com/locate/procedia

25th International Conference on Production Research Manufacturing Innovation: Cyber Physical Manufacturing
August 9-14, 2019 | Chicago, Illinois (USA)

Optimizing Mean and Variance of Multiresponse in a Multistage Manufacturing Process Using a Patient Rule Induction Method

Dong-Hee Lee[a]*, Kwang-Jae Kim[b]

[a]*Division of Interdisciplinary Industrial Studies, Hanyang University, 222 Wangsimni-Ro Seongdong-Gu, Seoul 04763, Korea*
[b]*Department of Industrial and Management Engineering, Pohang University of Science and Technology, 77 Cheongam-Ro Nam-Gu Pohang, Gyeongbuk 37673, Korea*

Abstract

Most manufacturing industries produce products through a series of sequential stages, known as a multistage process. In a multistage process, each stage is affected by its preceding stage, at the same time, it affects its following stage. Also, each stage often includes several response variables to be optimized. In this paper, we attempt to optimize the several response variables of the multistage process simultaneously considering the relationships among the stages. For this purpose, we use a particular data mining method, called a patient rule induction method. Because the relationships among the stages are often complicated, using a data mining method is a good approach for analyzing the relationships. According to the procedure of the patient rule induction method, the proposed method searches for an optimal setting of input variables directly from operational data at which mean and variance of the several response variables of the multistage process are optimized. The proposed method is explained by a step-by-step procedure using a steel manufacturing process example.

Keywords: *multiresponse optimization; multistage process; patient rule induction method; process optimization*

* Corresponding author. Tel.: +82-2-2220-2363; fax: +82-2-2220-2363.
E-mail address: dh@hanyang.ac.kr

2351-9789 © 2019 The Authors. Published by Elsevier Ltd.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0/)
Peer-review under responsibility of the scientific committee of the ICPR25 International Scientific & Advisory and Organizing committee members
10.1016/j.promfg.2020.01.433

1. Introduction

Response surface methodology (RSM) studies the relationship between a response variable (i.e., quality characteristic) and input variables (product or process variables) expected to influence the response. Its eventual purpose is to find a setting of input variables which optimize the response [1].

In most cases, considering several response variables is a common event in product and process development process. This kind of problem is called a multiresponse problem. A typical approach for solving the multiresponse problems is constructing empirical models for the responses under the RSM framework. This approach is called as multiresponse surface optimization (MRSO). Experiments are conducted by choosing various levels of the input variables, running the experiments, collecting the resulting responses and fitting models to relate the responses and the input variables [2]. Then, the setting of the input variables is obtained by analyzing the fitted models. This approach heavily relies on the empirical models so it works well only when the empirical models are fitted reasonably well.

However, it has been reported that the empirical models are often not fitted well to operational data from manufacturing lines whose volume is large [3]. Recently, many manufacturing companies can have the operational data from the manufacturing lines due to the development of information technologies such as IoT (internet of things) technology and big data analytics. When dealing with the operational data from manufacturing lines, a data mining approach is a good alternative.

The operational data from the manufacturing lines are often gathered in a multistage process which is a common structure in manufacturing lines. Figure 1 shows a typical multistage manufacturing process consisting of K sequential stages. The rectangles and circles represent the stages and inspection stations, respectively. The response of a semi-finished product is measured at the inspection stations after each stage. $\mathbf{x}_k$ and $\mathbf{y}_k$ are vectors of input variables and response variables, respectively, at the kth stage.

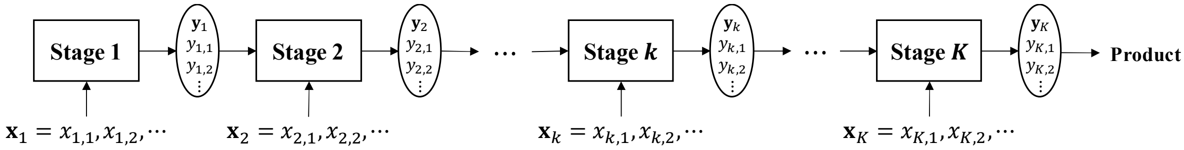

Fig. 1. Diagram of a multistage manufacturing process with multiple response variables.

The multistage process has two properties that are often encountered in practice. First, a result of the preceding stage affects the following stage. That is, the performance of Stage k is determined not only by $\mathbf{x}_k$, but also by $\mathbf{y}_{k-1}$. This relationship between stages is one of the most important characteristics of the multistage manufacturing process [4]. Secondly, it is common to have multiple responses at each stage which means that each stage has a multiresponse problem. The multiresponse problems in multistage process are not separated problems but related as mentioned in the first property. In such a case, the multiresponse problems should be solved simultaneously considering relationships between the stages. A large volume of process operational data is useful for solving the multiresponse problem of the multistage process, thus, using data mining technique is good approach for optimizing the multiresponse problem in multistage process.

One attractive data mining approach to process optimization is using a patient rule induction method (PRIM) [5]. PRIM searches a set of subregions of the input variable space for which the performance of the response is considerably better than that of the entire input domain [6]. In this paper, we propose a PRIM-based approach for solving the multiresponse problem in the multistage process. The proposed method attempts to optimize both of mean and variance of multiple responses in multistage process. Most of the existing multiresponse optimization methods assume that the variability of response is stable; thus, they focus mainly on the optimization of the mean of the multiresponse. However, this stable variability assumption often does not apply in most of practical situations; thus, the quality of the product or process can be severely damaged due to the high variability of multiple responses.

The rest of this paper is organized as follows. In Section 2, PRIM is reviewed. In Section 3, the proposed method based on PRIM is presented and illustrated with a case study. In Section 4, concluding remarks are given.

2. Patient rule induction method

The patient rule induction method (PRIM) has been successfully applied for process optimisation despite its recent emergence [3][4][6][7]. PRIM searches a set of subregions of the input variable space within which the performance of the response is considerably better than that of the entire input domain [6]. Here, a larger response is considered for illustration purposes. For the I input variables, $x_1, x_2, \ldots, x_I$, an I-dimensional box B is defined as the product:

$$B = s_1 \times s_2 \times \cdots \times s_i \times \cdots \times s_I.$$

Here, s_i is a sub-range of the ith input variable x_i, denoted by $s_i = (l_i, u_i)$, where l_i and u_i are the lower and upper limits, respectively.

Suppose that we have N observations denoted by $\{(y_n, \mathbf{x}_n), n = 1, 2, \ldots, N\}$, where y_n and $\mathbf{x}_n$ are values of the response and the input variables of the nth observation, respectively. $\mathbf{x}_n$ is an I-dimensional vector denoted by $\mathbf{x}_n = (x_{n1}, x_{n2}, \ldots, x_{nI})$. When $\mathbf{x}_n$ is located in box B (i.e., $l_1 \leq x_{n1} \leq u_1, l_2 \leq x_{n2} \leq u_2, \ldots, l_I \leq x_{nI} \leq u_I$), it is denoted by $\mathbf{x}_n \in B$.

Given a box B and observations $\{(y_n, \mathbf{x}_n), n = 1, 2, \ldots, N\}$, there are two statistics that describe the properties of box B. The first one is the support β_B, which denotes the proportion of the observations located in box B.

$$\beta_B = \frac{n_B}{N},$$

where n_B denotes the number of observations that are located inside box B. The second statistic is the box objective Obj_B, which is the mean value of the responses in box B.

$$Obj_B = \bar{y}_B = \frac{1}{n_B} \sum_{\mathrm{x}_n \in B} y_n.$$

The procedure for PRIM is presented below. Given observations $\{(y_n, \mathbf{x}_n), n = 1, 2, \ldots, N\}$, an initial box B_0 is formed by defining $l_i = \min_{n=1,2,\ldots,N} x_{ni}$ and $u_i = \max_{n=1,2,\ldots,N} x_{ni}$ for $i = 1, 2, \ldots, I$. From B_0, $2I$ candidate boxes, denoted by $\{C_{01-}, C_{01+}, C_{02-}, C_{02+}, \ldots, C_{0i-}, C_{0i+}, \ldots, C_{0I-}, C_{0I+}\}$, are created. The candidate boxes C_{0i-} and C_{0i+}, where $i = 1, 2, \ldots, I$, are obtained by peeling $100\alpha\%$ of the observations in box B_0 with respect to x_i. The parameter α determines the peeling rate, and its value is typically set between 0.05 and 0.1.

Once the $2I$ candidate boxes are obtained, PRIM chooses the one that has the largest box objective among the candidate boxes; this box becomes B_1. If the support of B_1 is greater than a stopping parameter β_0, then PRIM generates the next box B_2. This iterative process continues until the support of the box is less than the predetermined value, β_0.

3. The proposed method with a case study

3.1. Procedure of the proposed method

The proposed method, which we refer to as a robust multistage multiresponse PRIM (multistage MR-PRIM), is presented. The proposed method attempts to minimize "variability" of the response variables of the multistage process, thus, we call the proposed method "robust" multistage MR –PRIM. We assume that there are K stages ($k = 1,2,\ldots,K$), all of which need to be optimised. For the kth stage, there are I_k input variables, denoted by $x_{k1}, x_{k2}, \ldots, x_{ki}, \ldots, x_{kI_k}$ and there are J_k response variables, denoted by $y_{k1}, y_{k2}, \ldots, y_{kj}, \ldots, y_{kJ_k}$. Figure 2 shows the five-step procedure of the proposed method. Steps 3 and 4 iterate until the stopping criterion is satisfied. The details of each step are described below.

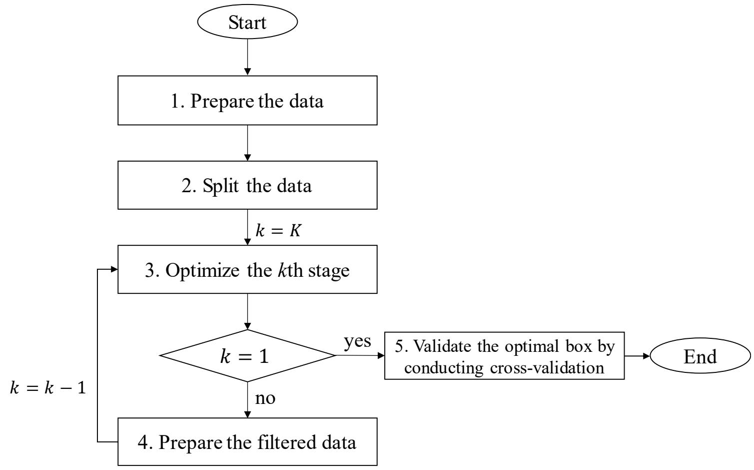

Fig. 2. Procedure of the proposed method

In Step 1, the entire dataset for K sequential stages are prepared. Suppose that we have N observations denoted by $\{(x_{nki}, y_{nkj})$ for $n = 1, 2, \ldots, N$; $k = 1, 2, \ldots, K$; $i = 1, 2, \ldots, I_k$; $j = 1, 2, \ldots, J_k\}$, where x_{nki} is the value of the ith input variable at the kth stage of the nth observation. Similarly, y_{nkj} is the value of the jth response variable at the kth stage of the nth observation. In Step 2, the dataset is split into learning and test sets for cross validation. The learning data set is used for Steps 3 and 4 to obtain the optimal condition for the input variables. The test et is used for Step 5 for the cross validation. In step 3, the kth stage is optimized by using a modified PRIM. This step begins with the final stage (i.e., $k = K$) and then the procedure continuous backward to the first stage sequentially. The modified PRIM consider variability of the response variables by defining the box objective as follows.

$$Obj_{B_k} = (d_{\mu_{k1}} \times d_{\mu_{k2}} \times \cdots \times d_{\mu_{kJ_k}} \times d_{\sigma_{k1}} \times d_{\sigma_{k2}} \times \cdots \times d_{\sigma_{kJ_k}})^{\frac{1}{2J_k}} \quad (1)$$

Obj_{B_k} is the box objective of the kth stage and it is a geometric mean of $2J_k$ desirability function values. The desirability function can be viewed as a utility function, and it ranges from 0 to 1. When a response variable achieves a target value, the desirability function value becomes 1, and the desirability function value decreases as the response variable deviates from the target value. d_{μ_k}'s and d_{σ_k}'s represent the desirability function values with respect to mean and standard deviation of the J_k responses at the kth stage, respectively.

In Step 4, filtered data are prepared to optimise Stage k-1. It should be noted that the filtering criteria come from the optimisation results for Stage k. Because Stage k-1 is optimised based on the filtered data, the optimisation results for Stage k are ensured. Steps 3 and 4 are repeated until $k = 1$. In Step 5, to avoid over-fitting problems, cross validation is conducted. The box objective obtained from the learning dataset is recalculated with the test dataset

3.2. Optimisation of the steel manufacturing process problem

The proposed method is illustrated by an example, specifically the steel manufacturing process problem. The steel manufacturing process consists of four stages to produce reinforcing bars. The scope of this case study, shown as the dotted box in Figure 3, is limited to the steel making and continuous casting stages (i.e., a two-stage process).

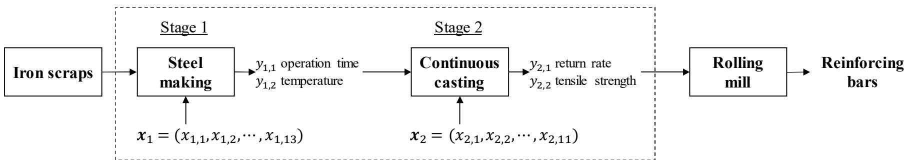

Fig 3. Steel manufacturing process in the case study.

In Stage 1, the iron scraps are melted to create molten iron. In the following Stage 2, the molten iron takes a solid shape, referred to as billets. The response variables of Stage 1 are the operating time ($y_{1,1}$, min) and molten iron temperature ($y_{1,2}$, °C). The operating time is a smaller-the-better (STB) type variable and the molten iron temperature is a nominal-the-better (NTB) type variable. If the molten iron temperature is too high or too low, the equipment can be damaged or billets are solidified prematurely. Thus, it is important to meet the target of the temperature. The target value of $y_{1,2}$ is 1580°C.

The response variables of Stage 2 are the return rate ($y_{2,1}$, %) and tensile strength ($y_{2,2}$, MPa). The return rate and tensile strength are larger-the-better (LTB) type variables. The return rate, an important performance measure in the steel industry, is defined as the ratio of the amount of iron scraps used in the steel-making stage to the number of billets produced in the continuous casting stage [4]. As illustrated with Figure 3, the number of input variables for each stage is 13 and 11, respectively. The purpose of this case study is to simultaneously optimise the means of multiple responses for each stage, while optimally screening that of the preceding stages. Each step is illustrated as follows.

- Step 1. Prepare the data

We have a total of 5609 operational data denoted by $\{(x_{nki}, y_{nkj})$ for $n = 1, 2, \ldots, 5609$; $k = 1, 2$; $i = 1, 2, \ldots, I_k$; $j = 1, 2, \ldots, J_k\}$. Here, the numbers of input variables of the first and second stages, denoted by I_1 and I_2, are 13 and 11, respectively. The numbers of response variables of the first and second stages, denoted by J_1 and J_2, are two for both, as shown in Figure 3.

- Step 2. Split the data

We randomly split the entire data into learning and test sets with a ratio of 2:1. Table 1 shows the subranges of input variables in Stage 2 which are obtained from the learning dataset. Note that the two response variables of Stage 1, $y_{1,1}$ and $y_{1,2}$ are treated as input variables in Stage 2.

Table 1. Sub-ranges of input variables for optimizing Stage 2.

	$y_{1,1}$	$y_{1,2}$	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	$x_{2,5}$	$x_{2,6}$	$x_{2,7}$	$x_{2,8}$	$x_{2,9}$	$x_{2,10}$	$x_{2,11}$
Lower	36	1447	1	13	1200	0	0	0	0	0	0	0	11
Upper	426	1666	264	121	12100	600	500	1425	25	270	120	89	163

- Step 3. Optimise the *k*th stage

Since there are two stages in this example, the stage iteration counter k is set as k=2. Then, the optimisation procedure for Stage 2 is performed with the learning set only. We set the initial parameters $\alpha = 0.05$ and $\beta_0 = 0.05$, which are typical values. Because $K = 2$, two box objective functions for Stages 1 and 2 are defined as follows.

$$Obj_{B_1} = \left(d_{\mu_{11}} \times d_{\mu_{12}} \times d_{\sigma_{11}} \times d_{\sigma_{12}}\right)^{\frac{1}{4}},$$
$$Obj_{B_2} = \left(d_{\mu_{21}} \times d_{\mu_{22}} \times d_{\sigma_{21}} \times d_{\sigma_{22}}\right)^{\frac{1}{4}}.$$

Table 2 summarises the results of optimisation for Stage 2. Some of the optimum conditions for variables are not peeled off, which are denoted by N/A. As can be seen in Table 2, the results indicate that the optimum conditions are significantly tighter than the current level of sub-ranges shown in Table 1.

Table 2. Optimal box input variables for optimizing Stage 2.

	$y_{1,1}$	$y_{1,2}$	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	$x_{2,5}$	$x_{2,6}$	$x_{2,7}$	$x_{2,8}$	$x_{2,9}$	$x_{2,10}$	$x_{2,11}$
Lower	40	1554	3	22	2700	200	N/A	400	3	15	2	4	40
Upper	46	1600	26	31	3700	295	0	1000	13	150	28	34	59

The optimal box given in Table 2 corresponds to an optimal setting of the input variables of Stage 2. We investigated process performance at this optimal box. The performance are shown as four indexes, $\bar{y}_{2,1}$, $\bar{y}_{2,1}$, $s_{y_{1,1}}$ and $s_{y_{1,2}}$ which are sample means and standard deviations of two responses at Stage 2, respectively. Samples satisfying the optimal box are participated in calculating the sample means and standard deviation. The last column of Table 3 reports the four statistics. Current case in Table 3 shows mean and standard deviation values before optimization. Compared to the current case, the proposed method improves all of the four performance indexes.

Table 3. Result of optimization of Stage 2.

	Current	Existing method	Proposed method
$\bar{y}_{2,1}$ (LTB)	89.85	92.47	91.82
$\bar{y}_{2,2}$ (LTB)	249.86	256.25	255.09
$s_{y_{2,1}}$ (STB)	5.91	4.56	4.11
$s_{y_{2,2}}$ (STB)	29.07	28.17	27.30

- Step 4. Prepare the filtered data

The dataset for optimizing Stage 1 is prepared by filtering data. The filtered data should meet the optimal conditions that were previously optimised in Step 3. More specifically, the data where $y_{1,1}$ and $y_{1,2}$ do not meet the optimal subranges obtained in Step 3 (i.e., [40, 46] for $y_{1,1}$ and [1447, 1666] for $y_{1,2}$) are excluded from the learning set. After the filtering process, the procedure goes back to Step 3 for optimizing Stage 1.

- Step 5. Validate the optimal box by conducting cross-validation

Cross-validation is conducted to investigate whether the optimal box is over-fitted. For this purpose, the box objective of the optimal box is recalculated using the test set, and then compared with the box objective value obtained from the learning set. We found that the box objective values obtained from the learning and test datasets are very similar and thus, we conclude that the obtained box is not over-fitted to the learning dataset.

3.3. Comparison with other methods

In the proposed method, we consider variability of the response variables as well as mean of the response variables. In this regard, we call the proposed method as "robust" multistage MR-PRIM. As mentioned in Section 1, ensuring stable variability of the response variables is important the quality of the product or process can be severely damaged due to the high variability of response variables. In this section, we compare the proposed method with existing PRIM called multistage MR-PRIM. Main difference between the existing method and the proposed method is that the existing methods does not consider the variability perspective. The existing multistage MR-PRIM uses desirability functions as the box objective as follows.

$$Obj_{B_k} = (d_{\mu_{k1}} \times d_{\mu_{k2}} \times \cdots \times d_{\mu_{kJ_k}})^{\frac{1}{J_k}}$$

When compared to Equation (1), the box objective function of the existing multistage MR-PRIM does not include desirability functions for standard deviation of the response variables. The last two columns of Table 3 compares mean and standard deviation values between the proposed method and the existing multistage MR-PRIM method. The proposed method shows better standard deviation values (i.e., lower values for the standard deviation) than the existing multistage MR-PRIM while mean of the two response variables are slightly worse than the existing multistage MR-PRIM. This means that the proposed method improves the standard deviation of the two responses while it sacrifices the mean of the two responses compared to the existing method.

Similar to Table 3, Table 4 compares mean and standard deviation values between the proposed method and the existing method. As expected, both of the existing and proposed methods shows better results than the current case. The proposed method shows better values for $\bar{y}_{1,1}$, $s_{y_{1,1}}$ and $s_{y_{1,2}}$ than the existing method. This means that the proposed method improves the three index (i.e., $\bar{y}_{1,1}$, $s_{y_{1,1}}$ and $s_{y_{1,2}}$) by sacrificing only one index (i.e., $\bar{y}_{1,2}$) compared to the existing method.

Table 4. Result of optimization of Stage 1.

	Current	Existing method	Proposed method
$\bar{y}_{1,1}$ (STB)	46.56	41.91	41.72
$\bar{y}_{1,2}$ (NTB)	1575.64	1579.83	1576.34
$s_{y_{1,1}}$ (STB)	15.26	1.56	1.37
$s_{y_{1,2}}$ (STB)	15.31	11.10	10.56

4. Concluding remarks

Most manufacturing industries produce products through a series of sequential stages. This is called a multistage manufacturing process. Contrary to RSM based MRSO approaches, the proposed method does not build the empirical models. Instead, it directly searches the optimal setting of the input variables by gradually peeling a large volume of operational data. Thus, it is free from the modeling issues such as low capability of the empirical models, model uncertainty, prediction variability and so on.

In this paper, we focused on the characteristics for the multistage manufacturing process as follows. First, the results of the preceding stage affect the following stage. Second, it is common to have multiple responses at each stage. Thus, it is important to simultaneously optimize the multiresponse of the multistage process in practice. In this paper, we proposed a robust multistage MR-PRIM, which attempts to optimize the mean and variability of the multiple responses of a multistage manufacturing process, and showed its performance with a case example. In the case study, the proposed method showed an improved performance compared to the performance at current operating condition as shown in Tables 3 and 4. Especially, the proposed method reduced the variability (i.e., standard deviation) for the two response variables compared to the existing multistage MR-PRIM.

Acknowledgements

This work was supported under the framework of an international cooperation program managed by the National Research Foundation of Korea (FY2017, 2016K2A9A2A11938390).

References

[1] R. H. Myers, D. C. Montgomery, C. M. Anderson-Cook, Response surface methodology: process and product optimization using designed experiments, 2016, John Wiley & Sons.
[2] D. Xu, S. L. Albin, Robust optimization of experimentally derived objective functions, IIE Transactions, 35(2003), 793-802.
[3] M. Lee, K. Kim, MR-PRIM: Patient Rule Induction Method for Multiresponse Optimization. Quality Engineering, 20(2008), 232-242.
[4] D. Kwak, K. Kim, M. Lee, Multistage PRIM: patient rule induction method for optimisation of a multistage manufacturing process. International Journal of Production Research, 48(2010), 3461-3473.
[5] J. H. Friedman, N. I. Fisher, Bump hunting in high dimensional data. Statistics and Computing -LONDON-, 9(1999), 123–142.
[6] I. G. Chong, S. L. Albin, C. H. Jun, A data mining approach to process optimization without an explicit quality function, IIE Transactions, 39 (2007), 795-804.
[7] I. G. Chong, C. H. Jun, Flexible patient rule induction method for optimizing process variables in discrete type, Expert Systems with Applications, 344(2008), 3014-3020.