Material Ingredient Optimization Based on Design of Experiment and Neural Network with Artificial Sample Generation

Juan Chen, Quan Sun, Jing Feng, and Zhengqiang Pan

***Abstract*—Material ingredient optimization is favorably and widely studied by using design of experiment (DOE) and artificial neural network (ANN). But the nonlinear mapping relationship model trained by the insufficient DOE samples can always cause non-negligible errors. This paper suggests a method adding some artificial samples into the neural network training data to get a better material ingredient optimization .In this method, artificial sample generation is combined with dimensionality reduction and segmentation technique. A simulation showed at the end of this paper indicates that compared with the model learning only from real DOE data, the accuracy can be significantly improved by adding some artificial training samples.**

***Index Terms*—Material ingredient optimization, DOE, artificial sample generation, insufficient samples.**

I. Introduction

Proportion of raw materials is of great importance to a product's performance and lifetime. Many researchers [1], [2] have studied the method of combing design of experiment(DOE) and artificial neural network(ANN) together to get the relationship model between the material ingredients and the related performance indicators of a product, which can then help to get a best-valued product by re-setting the ingredient parameters.

DOE can effectively figure out whether a factor has significant effect on the indicators or not. But its limitation in insufficient samples make the related ANN model have a non-negligible difference to the real relationship model, forcing it less practical in scientific and industrial use. Many researchers have studied the small sample problems in mapping relationship models and their findings can be summarized as below: 1.cross-validation.In this way, the scarce samples are divided into two sets, one set is used to train the model and the other one is applied to test the model. Y. M. Du [3] studied the effect of V-fold cross validation in remodeling back-propagation neural network (BPNN). U. Naonori [4] compared the accuracy of leave-one-out cross validation method and bootstrap analysis in ANN with the conventional training method.2.Adding some samples. These methods consider generating some artificial samples by information diffusion and training the relationship model by both samples to get a more accurate result. T. I. Tsai [5] studied a non-linear function fitting problem with one-dimensional input and one-dimensional output by segmentation technique and artificial sample generation. C. F. Huang [6] utilized diffusion-neural-network to fill the information gaps caused by data incompleteness. It seems that adding some samples have a more positive effect in lessening the modeling errors compared with cross-validation, but all these proposed sample generation methods are only used in one-dimensional input set problem, which still cannot be applied to DOE data of material ingredients.

This paper combines dimensionality reduction and segmentation technique to originate some artificial samples, and use both the real DOE data and the artificial samples to train a BPNN to reduce the error of the mapping relationship model and to achieve better material ingredient optimization.

Manuscript received February 6, 2015; revised April 10, 2015. This work was supported by National Natural Science Foundation of China under Grant No. 61273041.

The authors are with the School of Information Systems and Management, National University of Defense Technology, Sanyi Ave, Changsha, China (e-mail: 18874771120@163.com, sunquan@nudt.edu.cn).

II. BPNN Trained by DOE Data and Artificial Samples

In a mapping relationship model, changes of a factor could cause the related indicator change according to a regular trend. Thus, if the trend how a factor affects the according indicator is observed, artificial data could be generated by following this trend to partly fill the information gap .Much more feasible samples then could be used to train the supposed BPNN, leading it much closer to the real relationship.

Consider a DOE problem with multi-dimensional factors and one indicator. We can recognize all the trend information of the indicator affected by each factor and generate samples accordingly. This is quite time-consuming and the artificial samples may only differ in a limited scale. On account of this problem, we apply dimensionality reduction to map a multi-dimensional space into a single-dimensional space and the DOE factors X could then be converted to a single factor Z . By generating some artificial information (Z^*, K^*) in the new space and converting them to its original space randomly, the artificial samples (X^*, K^*) can be generated easily and can suffer a large diversity.

This paper firstly utilizes factor sensitivity analysis and dimensionality reduction to transform this multi-dimensional dataset problem to a one-dimensional set issue. Then picture the relation curve between the new factor and the corresponding indicators and segment the curve into some ascending and descending sections. In each section, quantities of single-dimensional factor values and their according indicator values are originated. After that, reverse the new space to its original space and get the artificial samples. Then, the ANN can be trained by both the real

DOI: 10.7763/IJMMM.2015.V3.200

sample data and the artificial sample data to achieve a better approximate result .Lastly, optimize the material ingredients to get the best value of the performance indicator by using the estimated BPNN model.

A. Factor Sensitivity Analysis

In a DOE problem, different factor can have different impact on the output indicator. This difference can be described as sensitivity. A little change of the factor with strong sensitivity can lead to a significant diversity in the indicator. To make sure this little change can cause the same significant diversity after the dimensionality reduction, the synthetical one-dimensional factor should also change obviously. On the other hand, the same change of a factor with weaker sensitivity should make the synthetical one-dimensional factor change less. That is to say, factor with strong sensitivity should hold a bigger proportion in the dimensionality reduction model. Principal component analysis (PLA) and other conventional dimensionality transform methods are no more practical here. This research propose getting the sensitivity of each factor at each level by analyzing the DOE data and use them as the parameters in dimensionality reduction model.

Suppose a DOE with n factors, marked as $X=\{x_1, x_2, \cdots, x_n\}$, and 1 indicator Y. Only d real samples are available. Let l_j be the number of levels of the j th factor. At each level, the j th factor should be applied in s_j times, $s_j = d / l_j$. Let P_{jm} be the value of the j th factor at level $m, m=1,2,3,\cdots, l_j$. K_{jm} is the sum of the indicator values related to P_{jm} and $\overline{K}_{jm}$ is the mean value of the indicator values related to P_{jm}, then

$$\overline{K}_{jm} = \frac{K_{jm}}{s_j} = \frac{K_{jm} \times l_j}{d} \tag{1}$$

R_{jm} is the sensitivity of the j th factor at level m ,then R_{jm} can be calculated as below:

$$R_{jm} = \begin{cases} \dfrac{\overline{K}_{j(m+1)} - \overline{K}_{jm}}{P_{j(m+1)} - P_{jm}}, & if \quad m=1; \\ 0.5\times\left(\dfrac{\overline{K}_{j(m+1)} - \overline{K}_{jm}}{P_{j(m+1)} - P_{jm}} + \dfrac{\overline{K}_{jm} - \overline{K}_{j(m-1)}}{P_{jm} - P_{j(m-1)}}\right), & if \quad 1<m<l_j; \\ \dfrac{\overline{K}_{jm} - \overline{K}_{j(m-1)}}{P_{jm} - P_{j(m-1)}}, & if \quad m=l_j. \end{cases} \tag{2}$$

R_{jm} reflects the change of the indicator when P_{jm} changes. The bigger R_{jm} is, the bigger the weight is at the dimensionality reduction model.

B. Dimensionality Reduction

When you submit your final version, after your paper has been accepted, prepare it in two-column format, including figures and tables. To figure out the trend of the indicator and fill the data gap by artificial samples, here we use a model to reduce dimensionality:

$$Z = \frac{\sqrt{\sum_{j=1}^{n}\left(\sum_{m=1}^{l_j} SN(c_j - m)\times(P_{jm} - D_{jm})^2 \times R_{jm}^{\ 2}\right)}}{\min\left|R_{jm}\right|} \tag{3}$$

where, c_j is the level of the j th factor of a specific sample. D_{jm} is the present best value of P_{jm}. $SN(i)$ is used to judge whether the specific sample is conducted at P_{jm} or not $SN(i) = \begin{cases} 1, if & i=0 \\ 0. if & i \neq 0 \end{cases}$. Thus, this multi-dimensional dataset problem then can be changed to a one-dimensional set issue.

C. Segmentation

Segmentation is used to find the trend of the indicator, and artificial data can be generated by following this trend.

The procedure of segmentation is described as following:

Step 1: According to the $Z \sim Y$ relationship figure (shown by an example in Fig. 1), group data into several ascending and descending sections. For a section with only one dot, the former data before this data is also included in this section.

The example shown in Fig. 1 is segmented into 4 sections, which is listed as Table I.

Step 2: Get the simple linear regression in each section. For a problem with N sections, N regression functions will be obtained, and we marked them as $y = f_i(z), i=1,2,\cdots, N$.

Step 3: Set the extreme points of each section by this 4 rules:

The intersection point $(\hat{z}, \hat{y})$ of regression lines of two contiguous sections lies between these sections. For section 1 and section 2 of the example showed in Fig. 2, the intersection point $(366.1967, 422.8079)$ lies between these two sections (shown in Fig. 2 (a)).The Z value of the right extreme

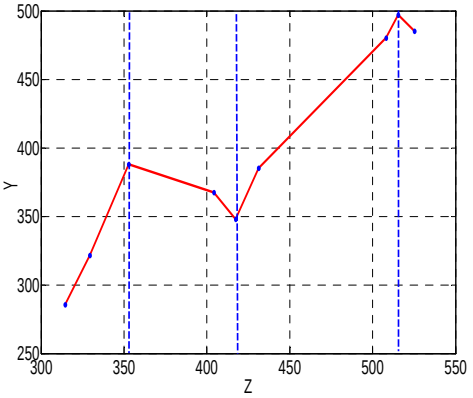

Fig. 1. $Z \sim Y$ relationship figure.

TABLE I: SEGMENTATION RESULT

section	trend	data
1	ascending	1,2,3
2	descending	4,5
3	ascending	6,7,8
4	descending	8,9

point of section 1, $R_{1,z}$, and the Z value of the left extreme point of section 2, $L_{2,z}$,can be set as $R_{1,z} = L_{2,z} = \hat{z} = 366.1967$ .The Y value $R_{1,Y}$ and $L_{2,Y}$ can be a random data between an interval of $\hat{y} \times (1 \pm \theta\%)$, $\theta\%$ is suggested as 5%.In this example, $R_{1,Y}$ and $L_{2,Y}$ can be any datum between [401.6675,443.9483], here we set $R_{1,Y} = L_{2,Y} = 420$.

The intersection point $(\hat{z}, \hat{y})$ is in or even before the preceding section. Set the right extreme point of the preceding section to be the right extreme point of the preceding section and the left extreme point of the latter section. For section 2 and section 3 of the example, the intersection point lies in section 2 (shown in Fig. 2 (b)).The last point of section 2 is chose to be the right extreme point of section 2 and the left extreme point of section 3.

The intersection point $(\hat{z}, \hat{y})$ is in or even after the latter section. Set the left extreme point of the latter section to be the right extreme point of the preceding section and the left extreme point of the latter section. For section 3 and section 4 of the example, the intersection point lies in section 4 (shown in Fig. 2(c)).The first point of section 4 is chose to be right extreme point of section 3 and the left extreme point of section 4.

The Z value of the left extreme point for the first section equals to the Z value of the first point and the Y value of this point is calculated by the regression function in section1. Similarly, the Z value of the right extreme point for the last section equals to the Z value of the last point, and the Y value of this point is calculated by the regression function in the last section.

Thus, all the available intervals for artificial sample generation in each section can be known.

D. Artificial Sample Generation

Generate some $Z \sim Y$ data at the available interval of each section. Restraints for artificial data generation are:

In an ascending section, if $Z_i < Z_j$, then $Y_i \le Y_j$;

In a descending section, if $Z_i < Z_j$, then $Y_i \ge Y_j$.

Here, we get a number of random values of Z and Y, sort them and match them into pairs. For data in an ascending section, match them in a consistent sequence. Conversely, in a descending section, match them in an adverse sequence.

For any pair of $Z \sim Y$, set any $n-1$ factors with any random values in their given intervals and calculate out the value of the left factor by using Eq. (1). If the value is acceptable, an appropriate sample with n factors is originated. If the value is not in the given interval, repeat setting the $n-1$ random factors until generating an eligible sample.

E. Neural Network Training

Combine the real data and the artificial data and train the BPNN. The topology of BPNN should be well-set in advance to get a more accurate relationship model.

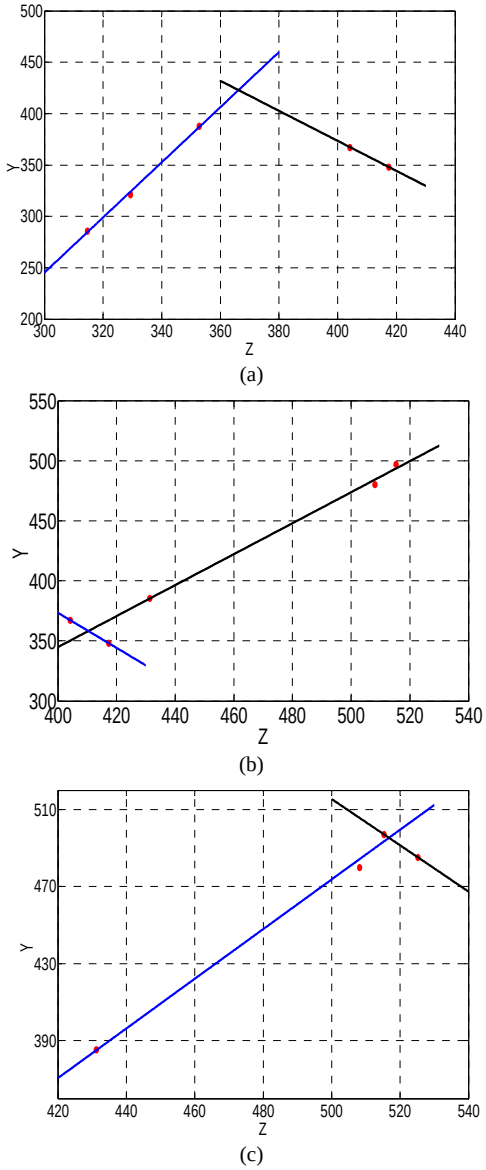

Fig. 2. Intersection points of each section.

The evaluation criterion used in this research is the standard deviation of error between the real value of indicators, y_i, and the output value of fitting model, $\tilde{y}_i$, that is:

$$SDE = \sqrt{\frac{\sum_{1}^{M} (\tilde{y}_i - y_i)^2}{M-1}} \tag{4}$$

F. Material Ingredient Optimization

After achieving the BPNN model, use nonlinear programing to set the appropriate value of each intergradient and to get products with better quality.

III. Simulation

There are many kinds of non-linear functions can be used as DOE simulation model. However if a function has continuous derivatives up to $(n+1)$ th order, then it can be expanded by Taylor Series as

$$f(x)=f(a)+f'(a)(x-a)+f''(a)\frac{(x-a)^2}{2!}+\cdots+f^{(n)}(a)\frac{(x-a)^n}{n!}+R_n \quad (5)$$

When n is bigger enough, $R_n \approx 0$,

$$f(x)\approx f(a)+f'(a)(x-a)+f''(a)\frac{(x-a)^2}{2!}+\cdots+f^{(n)}(a)\frac{(x-a)^n}{n!} \quad (6)$$

As a result, the polynormial functions can represent most of the non-linear functions. Here we use a polynormial function to explain this method in details and show its performance in reducing error.

$$y=\prod_{k=1}^{m}\sum_{j=0}^{m}c_{k,j}x_k^{\ j} \quad (7)$$

We use this multi-dimensional function as an example:

$$y=\prod_{k=1}^{m}\sum_{j=0}^{m}c_{k,j}(x_k-a_k)^j \quad (8)$$

where, $m=4$, $a=[5.05, 2.21, 1.26, 2.62]$

$$c=\begin{bmatrix}0.8 & 0.6 & 4.7 & 4.7 & 9.9\\ 2.7 & 1.0 & 4.0 & 2.2 & 7.5\\ 9.5 & 4.5 & 5.5 & 1.6 & 8.4\\ 7.1 & 0.2 & 3.2 & 2.2 & 4.3\end{bmatrix}$$

Product with better quality is indicated by a smaller y.

This is a model with 4 input factors and 1 output indicator. And the feasible value of each factor should be chose in the given intervals listed in Table II.

Here we suppose each factor with 3 levels. The best DOE table for a 4 factors with 3 levels is $L_9(3^4)$. Then choose the specific value of each factor at each level and calculate out the according indicators. Table III is the result.

Besides that, another 1000 random samples were used as testing data.

The sensitivity of each factor at each level is showed in Table IV.

TABLE II: INTERVALS OF FACTORS

factor	X_1	X_2	X_3	X_4
Interval	4.15~5.15	1.6~2.5	1.1~1.6	2.8~3.5

TABLE III: DOE RESULT

No	X_1	X_2	X_3	X_4	Y
1	4.2	1.8	1.2	3.0	250.00
2	4.2	2.1	1.3	3.2	3060.96
3	4.2	2.4	1.4	3.4	12719.25
4	4.6	1.8	1.3	3.4	4610.84
5	4.6	2.1	1.4	3.0	1957.58
6	4.6	2.4	1.2	3.2	6643.42
7	5.0	1.8	1.4	3.2	3097.29
8	5.0	2.1	1.2	3.4	9231.07
9	5.0	2.4	1.3	3.0	4601.08

The integrated one-dimensional factor z is computed by using (3). The $Z \sim Y$ data can be divided into 3 sections, and we generate 100 dataset at each section. The real $Z \sim Y$ data and the artificial data are demonstrated in Fig. 3. After a dimensionality transforms, a group of 300 artificial samples is generated.

TABLE IV: FACTOR SENSTIVITY

	X_1	X_2	X_3	X_4
Level 1	-2348.6	6990.5	-12838.7	9988.4
Level 2	374.7	8892.0	2749.4	16460.4
Level 3	3098.0	10793.5	18337.5	22932.5

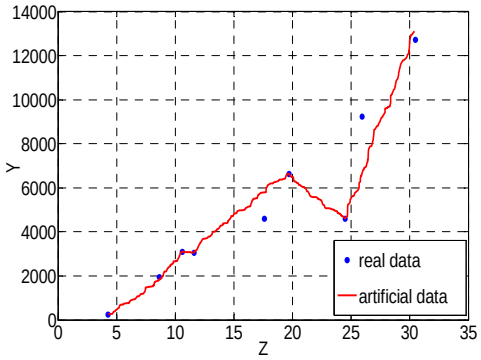

Fig. 3. $Z \sim Y$ Real data& artificial data.

TABLE V: SDE RESULT

without artificial samples	7221	9928	6427	9128	8347	Averag e:8210
with artificial samples	6569	4837	7327	5395	6550	Averag e:6136

The topology of neural network is set as '4-5-1'.Train and evaluate the BP network learning from both the real data and the artificial data and compare the result with the model learning only from real samples. Here, for each group of data, we train it 5 times and get the mean standard deviation of error. The computational results are listed in Table V.

Table V obviously illustrated that network with artificial samples show better accuracy compared with models trained without artificial samples, the improvement rate is as the following:

$$I=\frac{SDE_{without}-SDE_{with}}{SDE_{without}}\times 100\%=25.26\% \quad (9)$$

Other functions and other neural network topologies are also used to illustrate this method. It is obviously seen that the performance of this proposed method is much better than the network only learning from insufficient real samples and it could be better applied in material ingredient optimization.

IV. CONCLUSION

This paper studied the small sample problem in nonlinear mapping relationship models obtained from DOE data of material ingredients by adding some artificial samples to get a better accuracy. The artificial samples are generated according to factor sensitivity analysis, dimensionality reduction and segmentation technique. This method can successfully improve the BPNN model.

Future studies may pay attention to the relationship between the improvement rate and the parameters and rules used in this method. Different artificial sample generation

rules may have different impact on the accuracy of the model. For example, setting different ranges of the sections can leads to artificial samples with different characteristics and the model could be quite different. How to get a more accurate BPNN model is still under study.

REFERENCES

[1] G. P. Xiao and Z. K. Zhu, "Friction materials development by using DOE/RSM and artificial neural network," *Tribology International*, vol. 43, pp. 218-227, 2010.

[2] S. Mehdi, A. Hossein, and H. Mahtab, "Studing the abrasion behavior of rubbery materials with combined design of experiment-artificial neural network," *Chinese Journal of Polymer Science*, vol. 30, no. 4, pp. 520-529. 2012.

[3] Y. M. Du, Y. Li, and L. Z Chu, "Cave safety evaluation method study by small samples," *Enterprise Technology and Improvement*, in China, no. 22, 2009.

[4] U. Naonori and N. Ryohei, "Estimating expected error rates of neural network classifiers in small sample size situations: A comparison of cross-validation and bootstrap," *in Proc. IEEE International Conference on Neural Networks*, 1995, vol. 1, pp. 101-104.

[5] T. I. Tsai and D. C. Li, "Approximate modeling for high order non-linear functions using small samples set," *Expert Systems with Applications*, vol. 34, pp. 564-569, 2008.

[6] C. F. Huang and M. A Claudio, "Diffusion-neural-network for learning from small samples," *International Journal of Approximate Reasoning*, vol. 35, pp. 137-161, 2004.

Juan Chen is a master candidate in the School of Information Systems and Management, National University of Defense Technology, Changsha, China. Her major research interests are in the areas of reliability management, degradation modeling and health prognostics.

Quan Sun is a professor in the School of Information Systems and Management, National University of Defense Technology, Changsha, China. His research focuses on health prognostics, reliability analysis, degradation modeling and lifetime prediction.

Jing Feng is an associate professor in the School of Information Systems and Management, National University of Defense Technology, Changsha, China. Some of her interests include degradation models, lifetime analysis and management.

Zhengqiang Pan is a lecturer in the School of Information Systems and Management, National University of Defense Technology, Changsha, China. His major interests include degradation modeling and degradation experiment design.