Journal of University of Science and Technology Beijing
Volume 10, Number 1, February 2003, Page 69

A systematic method based on statistical pattern recognition for estimating product quality on-line

Guang Li, Huade Li, Shaoyuan Sun, and Zhengguang Xu

Information Engineering School, University of Science and Technology Beijing, Beijing 100083, China

(Received 2002-04-20)

Abstract: To avoid the complexity of building mechanistic models by studying the inner nature of the object, a systematic method based on statistical pattern recognition is developed in order to estimate the product quality on-line. The mapping relationship between a feature space and a product quality space can be built by using regression analysis, and in applying clustering analysis the product quality space can be partitioned automatically. Eventually, estimating product quality on-line can be accomplished by sorting the mapped data in the partitioned quality space. A concrete problem is proposed which has a relatively small ratio of training data to input variables. By implementing the method mentioned above, a satisfying result has been achieved. Furthermore, the further question about choosing suitable mapping methods is briefly discussed.

Key words: pattern recognition; regression analysis; clustering analysis; ISODATA algorithm; sorting algorithm

In modern manufacturing industries, it is usually necessary to estimate the quality of products on-line, because process engineers often use this information to ensure proper operation of plants. However, in some extremely complex manufacturing processes that involve a lot of intricate chemical and physical reactions, the product quality indices cannot be measured in real time. This is either due to the complicated and slow measurement techniques used or to the inability to measure the quality indices until the final product is formulated and used. On the other hand, plants are operated at desired conditions by setting and regulating process variables (such as pressures, temperatures, and flow rates), which can be measured readily on-line. These readily measurable variables are process variables in estimating product quality [1]. Therefore, an important problem proposed is how to use readily measurable variables to represent the quality indices not easily measured in real time.

1 Theoretical basis

In this paper, a 3-step method based on statistical pattern recognition is developed to solve the problem mentioned above.

The 3 steps are:

(1) Find the mapping relationship between the feature vector space and the quality vector space by using regression analysis.

(2) Partition the quality vector space by using clustering analysis.

(3) Project the feature vector to the quality vector space and locate the sites of the mapped data (on-line estimation of quality).

The procedure of implementing this method is illustrated in **figure 1**.

In the following sections, the theoretical basis used in these 3 steps is introduced separately, and a concrete instance is given to corroborate the feasibility of this method.

1.1 Regression analysis—the establishment of a mapping relationship

Regression analysis, an important branch of modern applied statistics, has been developed in Artificial Intelligence to analyze the regularity of different kinds of variables.

Regression analysis aims to find the statistical relation between one response and one or more independent variables. This relation can be expressed as:

$$Y=f(X)+\varepsilon \tag{1}$$

where Y is called a dependent variable vector; X is called an independent variable vector; ε is called a random disterburnce variable vector.

In different situations, the form of $f(x)$ can be expressed differently and the regression model would vary accordingly [2,3].

Corresponding author: Guang Li **E-mail:**liguang78@hotmail.com

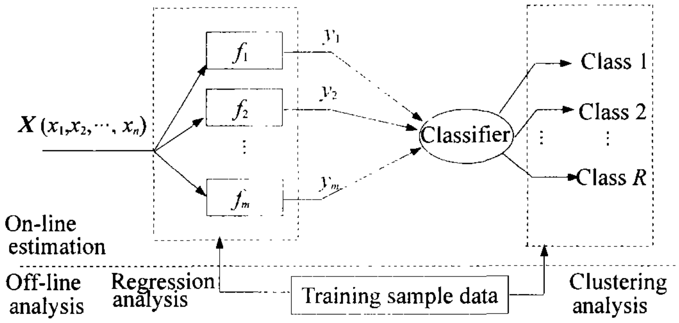

Figure 1 Quality estimation procedure, X (x_1, x_2, $\cdots$, x_n) —an input variable with n real-time measurable dimensions; f_1, f_2, $\cdots$, f_m—m mapping functions; (y_1, y_2, $\cdots$, y_m) —an output variable with m dimensions describing quality features; class1, $\cdots$, class R—R quality classes.

An elementary multiple linear regression model can be expressed as:

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k+\varepsilon \qquad (2)$$

where the dependent variable y is related to k regressor variables x_i, i=1,2,$\cdots k$. Models that are more complex in structure than equation (2) may still be analyzed by multiple linear regression technique. For example, consider the cubic polynomial model with one regressor variable, an interaction between two variables and a variable transformed by a triangular function:

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_1^3+\beta_4x_1x_2+\sin x_1+\varepsilon \qquad (3)$$

Let $x_3=x_1^3$, $x_4=x_1x_2$, and $x_5=\sin x_1$, then equation (3) can be written as:

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\varepsilon \qquad (4)$$

which is still a linear regression model defined by its coefficients β. By this transformation, the adaptability of the models is improved to approximate many complex statistical relations in real problems.

After a model is built, the next task is to estimate the coefficients. Least Squares Estimation of the Parameters is one of the efficient methods.

A sample regression equation can be expressed in matrix notation as:

$$Y=X\hat{\beta}+E \qquad (5)$$

Find the vector of least squares estimators, $\hat{\beta}$, that minimizes:

$$E'E=(Y-X\hat{\beta})'(Y-X\hat{\beta})=Y'Y-2\beta'X'Y+\hat{\beta}'X'X\hat{\beta} \qquad (6)$$

The least squares estimator $\hat{\beta}$ is the solution for $\hat{\beta}$ in the equation:

$$\frac{\partial(E'E)}{\partial\hat{\beta}}=-2X'Y+2X'X\hat{\beta} \qquad (7)$$

Then, the least squares estimator $\hat{\beta}$ is

$$\hat{\beta}=(X'X)^{-1}X'Y \qquad (8)$$

After the coefficient parameters have been acquired, certain tests of hypotheses about the model parameters are useful in measuring model adequacy. The test for significance of the regression is a test of the regressor variable y, and a subset of the regressor variables x_1, x_2, $\cdots$, x_k.

The appropriate hypotheses are:

H_0: $\beta_1=\beta_2=\cdots=\beta_k=0$;

H_1: $\beta_j\neq0$, for at least one j.

Define the regression sum of squares as:

$$\mathrm{SSR}=\sum_{i=1}^{n}(\hat{y}_i-\overline{Y})^2 \qquad (9)$$

and the error sum of squares as:

$$\mathrm{SSE}=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \qquad (10)$$

Then the test statistic is:

$$F=\frac{\mathrm{SSR}/p}{\mathrm{SSE}/(n-p-1)}\sim F(p,n-p-1) \quad (p=k+1) \qquad (11)$$

If $F\leq F_\alpha$ $(p,n-p-1)$, then H_1 is rejected; if $F\geq F_\alpha$ $(p, n-p-1)$, then H_0 is rejected. (α is called significance level). When H_0 is rejected, it shows that the response is linearly related to independents.

1.2 Clustering analysis—the partition of the quality vector space

In most real-world problems, it is sometimes very difficult to give a standard that can fully reflect the population quality level. This is due to the great variance of parameters that can be used to describe product quality. In this case, clustering analysis method can be used. Clustering is an activity of finding abstractions of data which can be used in decision-making. There are a variety of clustering algorithms; However, the Interactive Self Organizing Data Analysis Technique Algorithm (*i.e.*, ISODATA) is one of the most efficient due to its unmatchable advantages that include: the ability to combine and divide classes automatically, and the ability to acquire an appropriate class number as well as the corresponding clustering center of each class.

Its major steps are:

(1) Choose initial parameters (clustering center, *etc.*).

(2) Compute the necessary values, such as distance, variance, average, *etc.*

(3) Combine or divide the classes resulted from the last time.

(4) Evaluate the fitness of the clustering result. If the result is convergent, end the program; else go to step (1) and try other initial values [4,5].

While implementing this program, the following issues shoule be concerned.

(1) The selection of the initial values.

At the beginning of ISODATA program, need to set 8 values which define the number of initial classes, initial clustering centers, the least distance between every two clustering centers, the least number of sample data in each class, *etc.* The selection of these values influences the ultimate result greatly and the proper selection depends not only on the full understanding of the object properties but also on the amount of experiments on data.

(2) The proper use of the unit of each dimension.

Because the similarity between any two vectors in multidimensional space is determined by Euclid distance, the proper use of the unit of each dimension is also an important factor influencing final results.

(3) The mass value.

Since the contribution of each dimension to similarity is different, a mass value should be added to each dimension.

For example, Euclid distance should be written as:

$$D(x,y)=\|x-y\|=\sqrt{\sum_{i=1}^{n}\omega_i|x_i-y_i|^2} \tag{12}$$

where n is the number of dimensions of feature space; ω_i is the mass value of the i-th dimension.

1.3 Sorting algorithm—on-line estimation of product quality

After clustering analysis and regression analysis on sample data, the mapping relationship between the two spaces and the clustering center of each class in quality space are obtained. Next based on this knowledge the product quality can be estimated by processing data collected on-line.

Suppose that the sample data are classified into R different classes: ω_1, ω_2, $\cdots$, ω_R with corresponding clustering centers: Z_1, Z_2, $\cdots$, Z_R; then the estimation process follows 2 steps:

(1) Project the on-line data to the quality feature space as estimated quality value by linear Regression Equation.

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \rightarrow \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_m \end{bmatrix}$$

where n is the number of feature vector dimensions; m is the number of quality vector dimensions; $\hat{y}_i=f_i(x_1, x_2, \cdots, x_n)$, $i=1, 2, \cdots, m$.

For each regression formula f_i choosed, the following procedure should be followed: establish a series of different equations; estimate the coefficient values; test the feasibility of each equation (especially F test), and finally choose the superior equation according to tests as the mapping formula.

(2) Put the mapped value into Euclid distance formula:

If $\forall j=1, 2, 3, \cdots, R$ ($d_i<d_j$ and $j\neq i$), then $y\in\omega_i$.

Estimating product quality on-line is accomplished in this step.

2 Experimental

Agglomeration is an important production process before ironmaking. It's necessary to solve the problem of how to estimate the quality of the agglomerate [6,7]. Since the agglomerating process does not belong to Newton Mechanics Systems, the complexity of the process makes it impossible to solve the problem by the method of classical control theory, fuzzy control, expert system, *etc.*

Before use the method given above, some preliminary work must be done to obtain the needed data. The raw data were extracted from the pictures recorded beforehand. Finally, for every group of sample data, acquire 9 features describing the on-line condition of the agglomerating process and 3 product quality variables, which have mapping relations with the 9 features.

2.1 Clustering analysis

37 groups of sample data (numbered 1-37) are processed by ISODATA program. Choose 5 initial clustering centers: 1, 5, 10, 19, and 31. Clustering results are showed in **table 1.**

Table 1 Clustering results

Class number	Clustering center			Sample points
	y_1	y_2	y_3	
1	8.825 0	1.315 0	77.375 0	2, 27, 29
2	12.129 5	1.349 5	77.975 9	1, 3, 5, 6, 8, 9, 10, 12, 14, 15, 17, 21, 23, 24, 25, 36
3	14.297 7	1.298 5	77.739 6	4, 7, 11, 13, 16, 18, 19, 20, 28, 31, 32, 33, 35, 37
4	12.930 0	1.630 0	64.800 0	22, 30, 34

2.2 Regression analysis

In order to obtain a superior model for the output of each dimension, a variety of regression models should be built and tests for each model should be conducted. And then, compare the tests of models and adopt the superior. After a series of experiments and comparisons, the following models proved to be more suitable than others for the output of each dimension y_i (i=1, 2, 3):

$$y_1=-0.001\,8x_1-0.023\,3x_2+0.047\,9x_3+0.002\,2x_4+6.731\,0x_5+0.002\,4x_6-0.000\,1x_7-0.002\,9x_8+0.790\,4\sin 2x_9;$$

$$y_2=0.000\,9x_1+0.002\,2\cos x_2-0.042\,3x_3-0.000\,3x_4-2.228\,5x_5-0.000\,5x_6+0.000\,5x_7-0.000\,1x_8+0.012\,6x_9;$$

$$y_3=0.021\,5x_1+0.034\,4x_2+0.099\,1x_3+0.009\,6x_4+18.001\,3x_5+0.019\,2x_6+0.005\,4x_7+0.003\,6\cos x_8+1.103\,9x_9.$$

Analysis of Variance and F test of the 3 Regression models are shown in **table 2**.

2.3 Classification

In order to testify the feasibility of this method, 6 groups of sample data are classified by two different ways. One is to sort the real quality parameters directly; the other is to sort the estimated quality parameters mapped from the feature space by regression function.

The result of sorting the mapped quality parameters:

NO.1: NULL;

NO.2: 1, 2, 3, 5, 6;

NO.3: 4;

NO.4: NULL.

The actual classification result should be:

NO.1: NULL;

NO.2: 1, 2, 3, 4, 5, 6;

NO.3: NULL;

NO.4: NULL.

The comparison between the two results show that 5 of 6 data are correctly classified. If define the rate of correct classification as the percent of the number of correctly classified sample data to the number of total sample data, then the correct classification rate of this experiment is 83.3%. Through a lot of experiments on different sample data, the correct rate is around 85%. These results prove the feasibility of this method.

2.4 Discussion

In this experiment, a relatively good result have been achieved, but this can not prove that this method can be effective enough for any problems. The reason mainly lies in step (1), *i.e.* regression analysis.

Basically, the variety of empirical modeling methods can be divided into two categories: Artificial Neural Network (ANN) and statistical modeling methods. ANN (such as Back Propagation Networks) often requires a large amount of training data to obtain an acceptable model for a given number of input variables, whereas statistical methods can perform equally well with a smaller ratio of training data to input variables by using Regression Analysis Method [8,9]. Therefore, to different concrete problems, different suitable mapping methods should be choosed accordingly.

In this experiment, the total number of training data is 37, while the number of input variables is 9. In this case, the ratio of training data to input variables is relatively small which shows that regression analysis is more applicable than ANN.

Table 2 Analysis of variance and *F* test

Regressin function	SSR	MSR	SSE	MSE	Standard deviation	*F* test
f_1	28.695 3	3.188 4	63.635 6	2.356 9	1.311 4	1.352 8
f_2	0.427 5	0.047 5	0.300 1	0.011 1	0.090 1	4.279 3
f_3	95.5.67	10.611 9	183.943	6.812 7	2.229 7	1.557 7

3 Conclusion

Based on statistical pattern recognition for estimating product quality on-line a systematic method is proposed in this paper. This method mainly consists of 3 steps that include: regression analysis, clustering and sorting. Regression analysis is used for building the mapping relationship between the feature vector space and quality vector space; clustering is used for partitioning the quality vector space; and sorting is for locating the sites of mapped data in the partitioned space. By implementing this method based on statistical pattern recognition, the complexity of building the mechamistic model is avoided in estimating produce qualities.

References

[1] Masoud Soroush, State and parameter estimations and their applications in process control [J], *Computers & Chemical Engineering,* 23(1998), p.229.

[2] Douglas Montgomery, George C. Runger, and NormaFairs Hubeie, *Engineering Statistics* [M], John Wiley&Sons Inc, 2001.

[3] J.L. Devore, *Probability and Statistics For Engineering and-the Sciences* [M], Brooks/cole, 2000.

[4] Z.L. Jin, *Pattern Recognition* [M] (in Chinese), Higher Education Press, Beijing, 1994.

[5] V.S.Ananthanarayana, M.Narasimha Murty, and D.K.Subramanian, Rapid and brief communication Efficient clustering of large data sets [J], *Pattern Recognition,* 34(2001), p.2561.

[6] S.L. Tian, Iztok Livk, and Dean Ilievski, The influence of crystalliser configuration on the accuracy and precision of gibbsite crystallization kinetics estimates [J], *Chemical Engineering Science*, 56(2001), p.2511.

[7] I.Livk, M.Gregorka, and C.Pohar, Model identification of batch crystallization Processes [J], *Computers Chem. Engng,* 19(1995), Suppl., p.241.

[8] R. B. Bhavik and Raja Chatterjee, Unification of neural and statistical methods as applied to materials structure-property mapping [J], *Alloys and Compouds*, 279(1998), p.39.

[9] R.B. Bhavik and Utomo Utojo, Unification of neural and statistical methods that combine inputs by linear projection [J], *Computers Chem. Engng,* 22(1998), No.12, p.1859.