Proceedings of the 5th WSEAS International Conference on Applied Computer Science, Hangzhou, China, April 16-18, 2006 (pp497-502)

Development of a Data Mining Methodology using Robust Design

Sangmun Shin, Myeonggil Choi, Youngsun Choi, Guo Yi
Department of System Management Engineering,
Inje University
Gimhae, Kyung-Nam 621-749
South Korea

Abstract: The data mining (DM) method is far more effective than any other method when a large number of input factors are considered on a process design procedure. This DM approach to a robust design problem has not been adequately addressed in the literature nor properly applied to industries. As a result, the primary objective of this paper is two-fold. First, we show how DM techniques can be effectively applied into a process design by proposing a correlation-based factor selection (CBFS) method. Second, we then show how DM results can be integrated into a robust design (RD) paradigm based on the selected significant factors.

Key-Words: Data mining, *Response surface methodology*, *Robust design, Correlation-based factor selection*, *Best first search*

1. Introduction

Data mining (DA) is a term coined to describe the process of sifting through large databases for interesting patterns and relationships. This field spans several disciplines such as Databases, machine learning, intelligent information systems, statistics and expert system. Two approaches that enable standard machine learning algorithms to be applied to large databases are factor selection and sampling. Both approaches reduce the size of the database—factor selection by identifying the most salient factors in the data; sampling by identifying representative examples. According to our applications, in this paper we draw attention on the former approach.

Factor selection is an integral step of data mining process to find an optimal subset of factors. The factor selection algorithms perform a search through the space of feature subsets [1]. In general, two categories of algorithms have been proposed to solve factor selection problem. The difference of these algorithms is whether or not the factor selection is done independently of the learning algorithm. The first category is filter approach that is independent of an learning algorithm and serves as a filter to sieve the irrelevant factors. The second category is wrapper approach that uses the induction algorithm itself as part of the function evaluating factor subset [2]. Because all filter methods use heuristics based on general characteristics of the data rather than a learning algorithm to evaluate the merit of factor subsets as wrapper methods do, therefore, filter methods are generally much faster than wrapper methods, and, as such, are more practical for use on data of high dimensionality. The CBFS method is classified in the filter methods.

Most DA methods associated with the factor selection reported in literature may obtain a number of factors associated with the interesting response without providing the detailed information, such as relationships between the input factor and response, statistical inferences, and analyses [3], [4], [5], [6].

To address this situation, we first develop an enhanced robust design (RD) procedure integrating a DM methodology in order to select significant factors. The DM method is far more effective than any other method when a large number of input factors are considered on a process design procedure. This DM approach to a robust design problem has not been

adequately addressed in the literature nor properly applied to industry. As a result, the main purpose of this paper is two-fold. First, we show how DM techniques can be effectively applied into a process design by proposing a correlation-based factor selection (CBFS) method. This method can evaluate the worth of a subset including input factors by considering the individual predictive ability of each factor along with the degree of redundancy between pairs of input factors. Second, we then show how DM results can be integrated into a robust design paradigm based on the selected significant factors from the DM method, and that the robust design procedure based on the CBFS method can efficiently find significant factors and their associated statistical inferences.

2. Problem Formulation

2.1 Correlation-Based Factor Selection (CBFS)

Correlation-Based Factor Selection (CBFS) is a filter algorithm that ranks subsets of input factors 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 of a quality characteristic but also uncorrelated with each other (Hall 1999). Among input factors, irrelevant factors should be ignored because they may have low correlation with the given response. Even though some selected factors are highly correlated with the specified response, redundant factors must be screened out because they are also 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 the instance space not already predicted by other factors. The evaluation function of the proposed subset is

$$EV_S = \frac{n\overline{\rho}_{FR}}{\sqrt{n + n(n-1)\overline{\rho}_{FF}}} \qquad (1)$$

where EV_S, $\overline{\rho}_{FR}$, and $\overline{\rho}_{FF}$ represents the heuristic evaluation value of a factor subset S containing n factors, the mean of factor-response correlation $(F \in S)$, and the mean of factor-factor inter-correlation, respectively. $\sqrt{n + n(n-1)\overline{\rho}_{FF}}$ and $n\overline{\rho}_{FR}$ 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 factor and the response, an evaluation of a criterion called symmetrical uncertainty introduced in Section 2.2 is essential [3].

2.2 Symmetrical Uncertainty

In order to consider symmetrical uncertainty as a criterion, we may consider entropy that is a measure of the uncertainty or unpredictability in a given data set. Assuming a uniform manner of the specified response and all other factors based on the factor-response correlation $\overline{\rho}_{FR}$ and factor-factor inter-correlation $\overline{\rho}_{FF}$ in Eq. (1), Entropy of the specified response *Y* is

$$H(Y) = -\sum_{y \in Y} p(y)\log_2(p(y)) \qquad (2)$$

where $p(y)$ represents the probability of y value. If the values Y in the data set are partitioned according to the values of a second factor X, and the entropy of Y with respect to the partitions induced by X is less than the entropy of Y prior to partitioning, then there is a relationship between factors Y and X. Conditional entropy of Y given X can further be formulated as follows:

$$H(Y \mid X) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y \mid x)\log_2(p(y \mid x)). \qquad (3)$$

Based on entropy $H(Y)$ and conditional entropy $H(Y \mid X)$, the amount of decreasing entropy of Y called the information gain which is a symmetrical measure reflects additional information about *Y* given *X* [4]. Information gain can then be derived as follows:

$$gain = H(Y) - H(Y \mid X) = H(X) - H(X \mid Y).$$

$$= H(Y) + H(X) - H(X,Y) \quad (4)$$

The symmetrical measure represents that the amount of information gained about *Y* after observing *X* is equal to the amount of information gained about *X* after observing *Y*. Symmetry is a desirable property for a measure of factor-factor inter-correlation or factor-response correlation. Unfortunately, information gain is not apt to factors with more values. In addition, $\overline{\rho}_{FR}$ and $\overline{\rho}_{FF}$ should be normalized to ensure they are comparable and have the same effect. Symmetrical uncertainty can minimize bias of information gain toward factors with more values and normalize its value to the range [0, 1]. The coefficient of symmetrical uncertainty can be calculated by

$$C_{SU} = 2.0 \times \left[\frac{gain}{H(Y) + H(X)} \right]. \quad (5)$$

2.3 Best First Search Method (BFS)

In much literature, finding a best subset is hardly achieved in many industrial situations by using an exhaustive enumeration method. In order to reduce the search spaces for evaluating the number of subsets, one of the most effective methods is the best first search (BFS) method (Quinlan, R. R., 1986) which is a heuristic search method to implement CBFS algorithm. This method is based on an advanced search strategy that allows backtracking along a search space path. If the path being explored begins to look less promising, the best first search can back-track to a more promising previous subset and continue searching from there. The procedure using the proposed BFS algorithm is given by the following steps:

Step 1. Begin with the OPEN list containing the start state, the CLOSE list empty, and BEST← start state (put start state to BEST).

Step 2. Let a subset, $\theta = arg \max EV_S(\text{subset})$, (get the state from OPEN with the highest evaluation EV_S).

Step 3. Remove *s* from OPEN and add to CLOSED.

Step 4. If $EV_S(\theta) \geq EV_S(BEST)$, then BEST $\leftarrow \theta$ (put θ to BEST).

Step 5. For each next subset ξ of θ that is not in the OPEN or CLOSED list, evaluate and add to OPEN.

Step 6. If BEST changed in the last set of expansions, go to step 2.

Step 7. Return BEST.

2.4 The proposed data mining procedure based on Correlation-Based Factor Selection

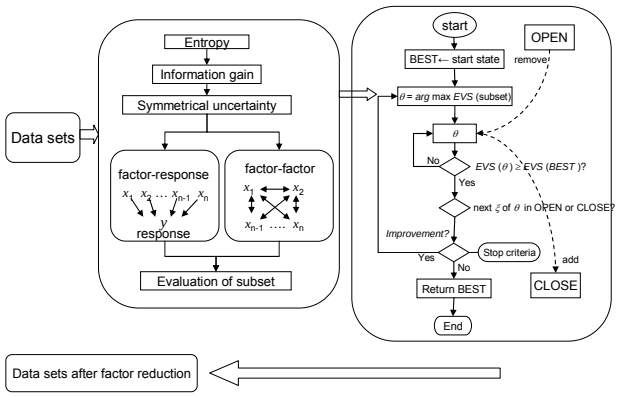
The evaluation function given in Eq. (1) is a fundamental element of CBFS to impose a specific ranking on factor subsets in the search spaces. In most cases, enumerating all possible factor subsets is astronomically time-consuming. In order to reduce the computational complexity, the BFS method is utilized to find a best subset. The BFS method can start with either no factor or all factors. The former search process moves forward through the search space adding a single factor into the result, and the latter search process moves backward through the search space deleting a single factor from the result. To prevent the BFS method from exploring the entire search space, a stopping criterion is imposed. The search process may terminate if five consecutive fully expanded subsets show no improvement over the current best subset. Figure 1 shows the entire process of the proposed CBFS algorithm. The CBFS method is used to calculate factor-response and factor-factor correlations using Eq (1). As a result, a factor subset with the highest evaluation value can be found.

3. Connection to robust design

Even though a data warehouse contains many factors including both controllable and uncontrollable factors which is known as noise factors. the proposed data mining method may provide significant factors associated with the given response. Based on the data mining solutions, a further analysis of the given solutions may also be an important part of a process design for applying the detailed and analyzed information to develop a process/product. In this

situation, RD principle can be utilized to provide statistical analyses and optimal factor settings for the selected factors associated with the given response by considering the effect of noise factors.

[Figure1. The proposed CBFS method]



3.1 Response surface methodology

Response surface methodology (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. RSM is a collection of mathematical and statistical techniques that are useful for the modeling and analysis of problems in which the response of interest is influenced by several variables and the objective is to optimize (either minimize or maximize) this response. For a comprehensive presentation of RSM, see [7], [8], [9], and [10].

Using responses for the interesting factor *y* and associated factors ***x***, the estimated response function for *y* is as follows:

$$\hat{y}(\mathbf{x}) = \hat{\alpha}_0 + \mathbf{x}^T\mathbf{a} + \mathbf{x}^T\mathbf{A}\mathbf{x} \tag{6}$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \mathrm{M} \\ x_k \end{bmatrix}, \quad \mathbf{a} = \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \mathrm{M} \\ \hat{\alpha}_k \end{bmatrix}$$

$$\mathbf{A} = \begin{bmatrix} \hat{\alpha}_{11} & \hat{\alpha}_{12}/2 & \Lambda & \hat{\alpha}_{1k}/2 \\ \hat{\alpha}_{12}/2 & \hat{\alpha}_{22} & \Lambda & \hat{\alpha}_{2k}/2 \\ \mathrm{M} & \mathrm{M} & \mathrm{O} & \mathrm{M} \\ \hat{\alpha}_{1k}/2 & \hat{\alpha}_{2k}/2 & \Lambda & \hat{\alpha}_{kk} \end{bmatrix},$$

x is the vector of the associated factors, and vector ***a*** and matrix ***A*** are the estimated regression coefficients for the interesting factor.

When using RSM, it is important to check that an estimated regression function is significant. An analysis of variance can be used to confirm that the regression function is indeed significant. The estimated regression functions can then be used for the optimization of the process parameters associated with the region of interest, such as the process mean and variance.

4. Problem Solution

4.1 Significant factor selection using data mining method

The data set comes from the daily measures of sensors in an urban wastewater treatment plant [4]. We select COND-S that is output conductivity of treated water as the response. Since the conductivity of water is an essential criterion for water purification, the lower the value of conductivity is, the purer the water is. One of the indispensable purposes of water treatment is to reduce the conductivity of water.

The data set contains 34 factors and 527 instances [4]. Among 34 factors, COND-S may include uncertain effects that are either irrelevant or redundant. If potential significant factors are selected by subjective opinions or experiences, it may often not include important factors on a factor selection process. Our objective is to find the most significant factors to the output response by short time consuming. In particular, during the two-step process of wastewater treatment, we want to make sure whether some input factors can affect the response factor significantly. The factors of the water treatment data set are shown in Table 1. Factor2-22 cover all input values measured during the process of two-step treatment, and factor 23-29 cover all output criterion values after two settlers treatment, and factor 30-38 cover the performance criterions.

[Table 1. The Water-Treatment Plant Data Set]

Q-E	ZN-E	PH-E	DBO-E	DQO-E	…	SED-D	COND-S
35023	3.5	7.9	205	588	...	0.4	2060
29156	2.5	7.7	206	451	...	0.3	1233
39246	2	7.8	172	506	...	0.6	1825
42393	0.7	7.9	189	478	...	0.4	1562
40923	3.5	7.6	146	329	...	0.2	1467
43830	1.5	7.8	177	512	...	0.4	1401
...	...	...	...	...	...	...	...

4.1 factor selection result

Table 2 shows the evaluation result of the numeric example calculated by the DM software named "Weka" [6]. The merit of best subset equals 0.92, the highest value among the calculated 371 subsets. The factor set F = {ZN-E, SED-D, COND-E, COND-P, SS-S, RD-DBO-P, DQO-S} is considered the best factor subset towards the response factor COND-S. Aside from the output factors including SS-S, RD-DBO-P and DQO-S, we get the best input factor subset BFS = {ZN-E, SED-D, COND-E, COND-P}. Among the BFS, COND-E represents the observation value of initial input conductivity to plants, therefore, it can hardly be controlled during the RD process. Consequently, we consider COND-E the noise factor, and consider other input factors among BFS the controlled factors.

4.2 Response surface methodology based on both control and noise factors

The data mining solution provides the four significant factors as ZH-E, COND-P, SED-D, and COND-E. Among these solutions, a primary input conductivity, COND-E, may often not be controlled in a water treatment process. For this reason, we regard COND-E as a noise factor incorporating the RD principle in order to achieve a robust process and the other factors as control factors. As shown in Figure 2, the regression is significant based on the results of the global F-test and its associated p-values. In addition, the response model has 85% R-sq, which implies the model may adequate to utilize as a response function.

[Table 2. Experiments result]

Selected Evaluator	The response attribute	COND-S
	Merit of best subset	0.92
	Selected attributes	ZN-E, SED-D, COND-E, COND-P,SS-S, RD-DBO-P, DQO-S
Search method	Search method	Best First
	Search Direction	forward
	Start set	no attributes
	Total number of subsets evaluated	371

5. Conclusion

In this paper, we developed an enhanced process design method by integrating DM method to select significant factors associated with the given response to an RD method to provide best factor settings. Based on the factor selection procedure including CFBS method and BFS heuristic search method on the DM stage, we quickly find important factors for water treatment process among a large set of data including many factors. When utilizing BFS method, the proposed CFBS method in its pure form is exhaustive, but the use of a stopping criterion makes the probability of searching the whole data set quickly. We then analyze the factor selection results using RD and RSM while incorporating a noise factor for uncontrollable one. Further, we will integrate more DM techniques and conduct other recent RD methods. For further studies, we may find the best factor settings using an RD optimization based on the analyzed RD results.

```
Estimated Regression Coefficients for y

Term          Coef     SE Coef       T        P

Constant     546.13    103.060     5.299     0.000

x1            -16.00      15.225    -1.051     0.294

x2             -0.43      0.493    -0.867     0.387

x3           174.80    113.587     1.539     0.125

z1             0.79      0.487     1.615     0.107

x1*x1         -0.63      0.580    -1.088     0.277
```

References:

[Figure 2. Outputs for response surface analyses]

[1] Allen, D., The relationship between variable selection and data augmentation and a method for prediction, *Technometrics*, Vol.16, 1974, pp. 125–127.

[2] Langley, P., Selection of relevant features in machine learning, *Proceedings of the AAAI Fall Symposium on Relevance*, AAAI Press, 1994, pp. 140-144.

[3] Press, W.H., Flannery, B.P., Teukolsky, S.A., and Vetterling, W.T. *Numerical Recipes in C.* Cambridge University Press, Cambridge, UK, 1988.

[4] Quinlan, R.R. Induction of decision trees, *Machine Learning*, Vol1, 1986, pp. 81-106.

[5] Gardner, M. and Bieker, J., Data mining solves tough semiconductor manufacturing problems, *Conference on Knowledge Discovery in Data Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2000, pp. 376-383.

[6] Witten, I.H. and Frank, E., *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, 2nd Edition, Morgan Kaufmann, San Mateo, CA, 2005.

[7] Box, G.E.P., Bisgaard, S., and Fung, C., An explanation and critique of Taguchi's contributions to quality engineering, *International Journal of Reliability Management*, Vol.4, 1988, pp. 123-131.

[8] Vining, G.G. and Myers, R.H., Combining Taguchi and response surface philosophies: A dual response approach, *Journal of Quality Technology* Vol22, 1990, pp. 38-45.

[9] Del Castillo, E. and Montgomery, D.C., A nonlinear programming solution to the dual response problem, *Journal of Quality Technology*, Vol25, 1993, pp. 199-204

[10] Shin, S. and Cho, B.R., Bais-specified robust design optimization and its analytical solutions, *Computers & Industrial Engineering*, Vol48, 2005, pp 129-140.