A Surrogate Variable-Based Data Mining Method Using CFS and RSM

Le Yang[1], Sangmun Shin[1]*, Yongsun Choi[1], Myeonggil Choi[1], Younghee Lee[2]
[1]Department of Systems Management & Engineering, Inje University
607 Obang-Dong, Gimhae, Gyungnam 621-749, South Korea
[2]Department of Industrial Management Engineering, Dong-A University
840 Handang-Dong, Saha, Busan 604-714, South Korea

Abstract: - In many scientific and engineering fields, there are a number of data sets uncontrollable and hard to handle because the nature of measurement of a performance variable may often be destructive or very expensive, which are known as sets of noise factors. Although these noise factors, which may not be controlled by manufacturing and cost reasons, are merged as a key problem of data mining (DM) and analysis, most DM methods may not discuss robustness of solutions either by considering noise factors or by incorporating specific statistical inferences. In order to address this problem, the primary objective of this paper is to propose a integrated approach， called surrogate variable-based data mining method (SVDM), which can conduct dimensionality reduction by exacting the significant factors from the row data sets by applying correlation-based feature selection (CFS). The proposed method then incorporates noise factor consideration to achieve robustness of an analysis by using the principle of surrogate variable. In addition, this proposed method is far more effective when a 100% inspection and a destructive characteristic/response are considered. Finally, response surface methodology (RSM), which is a statistical tool that is useful for modeling and analysis in situations where the response of interest is affected by several input factors, is used for further statistical analyses.

Key-Words: - Data mining, Surrogate variable, Correlation-based feature selection (CFS), Response surface methodology (RSM)

1 Introduction

The continuous improvement and application of the information system technology has become widely recognized by industry as critical in maintaining a competitive advantage in the marketplace. It is also recognized that the improvement and application activities are most efficient and cost-effective when implemented during an early process/product design stage. Real world data is often not perfect to conduct analysis and also suffers from uncontrollability that may impact interpretations of the data, models created form the data, and decisions made based on the data [1]. In many scientific and engineering fields, there are a number of data sets uncontrollable and hard to handle because the nature of measurement of a performance variable may often be destructive or very expensive, which are known as sets of noise factors. These noise factors, which may not be controlled by manufacturing and cost reasons, are merged as a key problem of data mining (DM) and analysis. It may be possible to increase the specificity of a DM algorithm if a number of the noise factors presented in a data set can be reduced. In other words, if some of the noise can be removed from a data set, the model can better learn the underlying relationship. Any noise-removing techniques that are applied to a source data set have to be duplicated on any run-time data during deployment, so the transformations to reduce noise have to be carried forward [2]. In order to effectively address these problems associated with noise factors, utilization of a surrogate variable approach may be an alternative method. When the nature of input factors is destructive or costly, it is possible to use a surrogate variable which is highly correlated with the input factors and less expensive to measure instead of using the original noise factors. For example, when design engineers are interested in measuring the strength of industrial glasses implemented a 100% inspection on a key factor, and all glasses need to be destroyed to measure the strength for testing purposes (i.e., the nature of the input factors are destructive), implementing a

surrogate variable may be more practical.

DM has emerged as one of the key features of many applications on computer science. Often used as a means for predicting the future directions, extracting the hidden limitations, and the specifications of a product/process, DM involves the use of data analysis tools to discover previously unknown, valid pattern and relationships from a large database. Most DM methods associated with the factor selection reported in literature may obtain a number of factors associated with the interesting response factor without providing the detailed information, such as relationships between the input factor and response, statistical inferences, and analyses ([3], [4], [5], and [6]). Su et. al [7] developed an integrated procedure combining a DM method and Taguchi methods. The factor selection algorithm performs a search through the space of feature subsets [8]. 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 [9]. Because most filter methods is based on a heuristic algorithm for general characteristics of the data rather than a learning algorithm to evaluate the merit of factor subsets as wrapper methods do, filter methods are generally much faster, and has more practical capabilities to utilize high dimensionality than wrapper methods.

Most DM methods may not discuss robustness of solutions either by considering noise factors or by incorporating specific statistical inferences. In order to address the problem related to noise factors in data mining methods, we proposed a integrated approach , called surrogate variable-based data mining method (SVDM) as shown in Fig. 1. The main purpose of this paper is four-fold. First, the proposed SVDM can conduct dimensionality reduction for a large number of data sets including many input factors and responses by exacting the significant factors from the row data sets by applying correlation-based feature selection (CFS). Second, the proposed method incorporates noise factor consideration to achieve robustness of an analysis by using the principle of surrogate variable. Third, this proposed method is far more effective when a 100% inspection and a destructive characteristic/response are considered. When the analysis results provide a noise factor as a significant factor, a surrogate variable, which is a redundant and highly correlated factor with the interesting response, can be used as a key factor for further analyses instead of using the noise factor. Finally, response surface methodology (RSM), which is a statistical tool that is useful for modeling and analysis in situations where the response of interest is affected by several input factors, is used for further statistical analyses.

This paper is organized as follows. In section 2, we conduct a generic literature review associated with the principles and applications of the surrogate variable. In Sections 3 and 4, we then present the proposed our integrated data mining method based on surrogate variable using the correlation-based feature selection (CFS) and response surface methodology (RSM), respectively. In section 5, we conclude the paper and future work.

2 Surrogate Variable

Surrogate variable technique is a sub branch of screening inspection. Nowadays, it is very popular to utilize screening inspection which can mainly be divided into two kinds of methods. Firstly, a performance variable can be measured directly. Secondly, a surrogate variable which is highly correlated with a performance variable can be measured. And sometimes, it is difficult to measure a performance variable directly. For instance, when we need a destructive inspection for a product or the charge tilizing a performance variable for a quality audit is too high, and if we also can find out a surrogate variable which is highly correlated with performance variable and the inspection cost is relatively low, we can put up a screening inspection which is utilizing a surrogate variable.

Generally, the nature of measuring or observations on a response (i.e., a dependent variable) may be exceptionally expensive and destructive or hard to obtain, forcing a reduction in the overall sample size used to fit the model.

In our approach, noise factors of significant factors, both in response and in input, are referred to the destructive or very expensive performance variables to measure. With the effort of CFS aforementioned, we can easily find the candidate surrogate variables from redundancy factors for every noise factor. Figure 2 shows us two cases of surrogate (k, i, n, m $\in$ Int). One is when one of the interesting responses yn have a characteristic of noise, while another interesting response ym is not only highly correlated to the noise one but controllable, the surrogate between y_n and y_m is becoming considerable. Another is when we focus on a specific interesting response y_k corresponding to the some input factors (x_k, x_i, ……, x_n) which the factor X_i is

noise, however, factor x_1 is neither noise and irrelevant to X_i ,nor corresponding to interesting response y_k. Then the x_1 will be the available surrogate variable candidate for x_i.

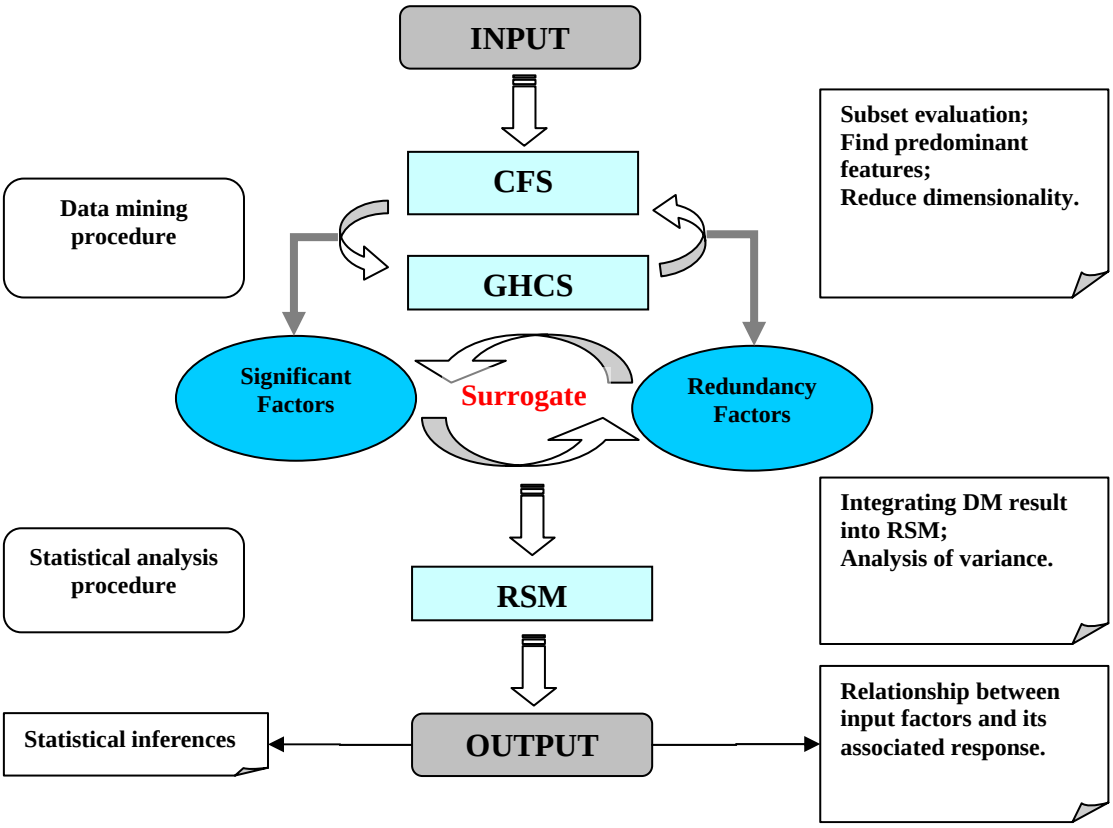

Fig. 1. Overview of the proposed integrated DM method: The first stage presents a procedure of significant feature selection using CFS method; surrogate the noise factors of significant features by its highly correlated redundancy features. And the second stage implies a specific statistical analysis using RSM based on the results of the previous stage.

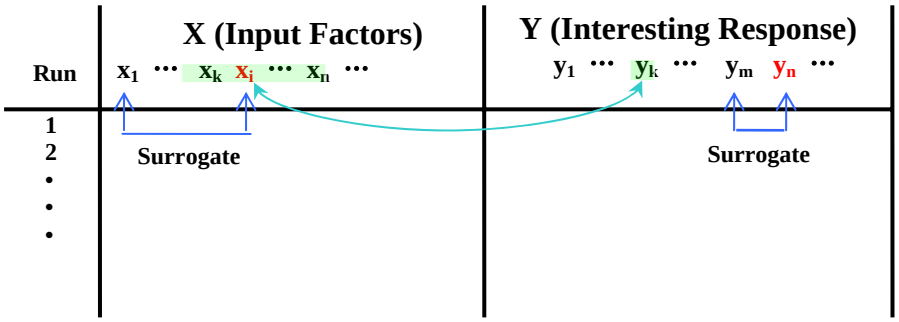

Fig. 2. Two Cases of surrogate variable

3 Data Mining Method

3.1 Correlation-based Feature Selection (CFS)

Correlation-based Feature Selection (CFS) 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 ([9] [10] [11]).

Among input factors, irrelevant factors should be ignored because they may have low correlation with the given response. Although 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

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

where, EVs, $\overline{\rho}_{FR}$, $\overline{\rho}_{FF}$ 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 [12].

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 features with more values and normalize its value to the range [0, 1]:

Symmetrical uncertainty =

$$2.0 * \left[\frac{gain}{H(Y)+H(X)} \right] \quad (2)$$

Where:

$$H(Y) = -\sum_{y \in Y} P(y)\log_2(P(y))$$

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

gain = H (Y) – H (Y|X) = H (X) – H (X|Y) = H (Y) + H (X) – H (X, Y)

and where H(Y), p(y), H(Y|X), and gain represent the entropy of the specified response Y, the probability of y value, the conditional entropy of Y given X, and the information gain that is a symmetrical measure reflects additional information about Y given X, respectively.

3.2Greedy Hill Climbing Search (GHCS) Algorithm

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 Greedy Hill Climbing Search (GHCS) method which is a heuristic search method to implement CBFS algorithm. These search strategies such as forward selection and backward elimination [13] are often applied to search the feature subset space in reasonable time. Although simple, these searches often yield sophisticated AI search strategies such as Best First search and Beam search [14].

Greedy hill climbing expands the current node and moves to the child with the highest evaluation. Nodes are expanded by applying search space operators to them (add or delete a single factor for forward selection and backward elimination respectively). The procedure using the proposed GHCS algorithm is given by the following steps:

Step 1. Let s ← start state.
Step 2. Expand s by applying search operators.
Step 3. Evaluate each child t of s.
Step 4. Let s' ← child t with highest evaluation e (t).
Step 5. If e(s') > e(s) then s ← s', go to step 2.
Step 6. Return s.

The evaluation function given in Eq. (1) is a fundamental element of CFS 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 GHCS method is utilized to find a subset with highest evaluation. The CFS 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. Figure 3 shows the overview of CFS.

4 Response Surface Methodology (RSM)

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 attributes. 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, Box et al. [12] and Shin and Cho [15] provide insightful comments on the current status and future direction of RSM. Using this method, the response function evaluation is given in our approach by

$$\hat{y}(x) = \hat{a}_0 + x^T a + x^T A x \tag{3}$$

Where

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad a = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix}, \quad \text{and}$$

$$A = \begin{bmatrix} \hat{\alpha}_{11} & \hat{\alpha}_{12}/2 & \cdots & \hat{\alpha}_{1k}/2 \\ \hat{\alpha}_{12}/2 & \hat{\alpha}_{22} & \cdots & \hat{\alpha}_{2k}/2 \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\alpha}_{1k}/2 & \hat{\alpha}_{2k}/2 & \cdots & \hat{\alpha}_{kk} \end{bmatrix}$$

Where xi terms are control factors, and the estimate of the α's in the function is estimated regression coefficients of the second-order fitted response function.

5 Conclusion

In this paper, we proposed a SVDM method by integrating a CFS method for finding significant factors, then utilize the principle of the surrogate variable which is a highly correlated redundant factor instead of using the noise factor. We then discussed a RSM method for further statistical analyses. The CFS method in its pure form is exhaustive, but the use of a stopping criterion makes the probability of searching the whole data set quickly. For further research, , the consideration of outliers of data using expectation maximization (EM) algorithm in order to achieve better precision of the proposed DM method can be an possible further research issue.

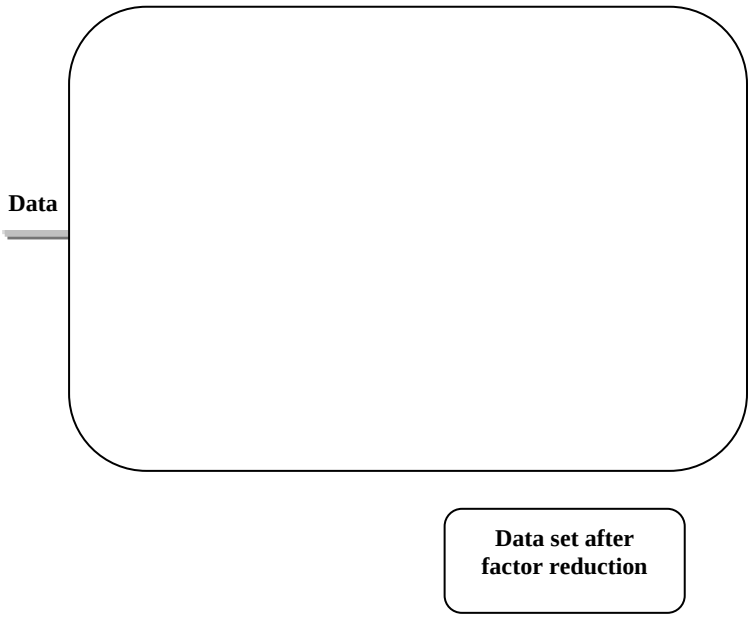

Fig. 3. Overview of CFS method

Reference:

[1] J. Kubica and A. Moore, Probablilistic Noise Identification and Data Cleaning, Proceedings of the Third IEEE International Conference on Data Mining , 2003, pp: 131-138.

[2] D. Pyle, Business Modeling and Data Mining, Morgan Kaufmann Publishers, 2003

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

[4] R.R. Quinlan, Induction of Decision Trees. Machine Learning. Vol. 1, Hingham, MA 1986, pp. 81-106.

[5] M. Gardner, and J. Bieker, 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] I.W.H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques. 2nd edn Morgan Kaufmann, 2005.

[7] C.T. Su, M.C Chen and H.L. Chan, Applying Neural Network and Scatter Search to Optimize Parameter Design with Dynamic Characteristics. Journal of the Operational Research Society, Vol. 56 , 2005, pp.1132-1140

[8] D. Allen, The Relationship between Variable Selection and Data Augmentation and a Method for Prediction, Technometrics. Vol. 16, 1974, pp. 125–127.

[9] P. Langley, Selection of Relevant Features in Machine Learning. Proceedings of the AAAI Fall Symposium on Relevance, AAAI Press 1994, pp. 140-144.

[10] J.W. Seifert, Data Mining: An Overview. CRS Report RL31798, 2004.

[11]Q. Xu, M.Kamel and M.M.A. Salama, Significance Test for Feature Subset Selection on Image Recognition. International Conference of Image Analysis and Recognition LNCS, Vol. 3211 , 2004, pp. 244-252.

[12]G.E.P. Box, S. Bisgaard and C. Fung,: An Explanation and Critique of Taguchi's Contribution to Quality Engineering. International Journal of Reliability Management. Vol. 4, 1998, pp.123-131.

[13]J.Kittler, Feature set search algorithms, in C.H Chen(ed) Pattern Recognition and Singal Processing, Sijhoff and Noordhoff, the Netherlands, 1978.

[14]E.Rich and K.Knight, Artificial Intelligence, McGraw-Hill,1991.

[15] S. Shin and B.R. Cho, Bias-specified robust design optimization and its analytical solutions. Computerr & Industrial Engineering. Vol. 48, 2005, pp.129-140.