4

Robust Data Mining: An Integrated Approach

Sangmun Shin, Le Yang, Kyungjin Park and Yongsun Choi
Department of Systems Management & Engineering, Inje University
Korea

1. Introduction

The continuous improvement and application of information system technologies have become widely recognized by the industry as critical for maintaining a competitive advantage in the marketplace (Shin et al., 2006). It is also recognized that improvement and application activities are the most efficient and cost-effective when implemented during an early process/product design stage. Data mining (DM) has emerged as one of the key features of many applications in computer science. Often used as a means for predicting the future directions and extracting hidden limitations and specifications of a product/process, DM involves the use of data analysis (DA) tools to discover previously unknown and valid patterns and relationships from a large database. Most DM methods for factor selection reported in literature may yield a number of factors associated with interesting response factors without providing detailed information, such as relationships between the input factor and response, statistical inferences, and analyses (Yang et al., 2007; Witten & Frank, 2005). Based on this, Gardner and Bieker (Gardner & Bieker, 2000) suggested an alternative DA approach toward resolving semiconductor manufacturing problems in order to determine the significant factors. Furthermore, Su et al. (Su et al., 2005) developed an integrated procedure combining a DM method and Taguchi methods.

DA is a term coined to describe the process of sifting through large databases for discovering interesting patterns and relationships. This field spans several disciplines such as databases, machine learning, intelligent information systems, statistics, and expert systems. Two approaches that enable the application of standard machine learning algorithms to large databases are factor selection and sampling. Factor selection is known to be an effective method for reducing dimensionality, removing irrelevant and redundant data, increasing mining accuracy, and improving result comprehensibility (Yu & Liu, 2003). Consequently, factor selection has been a fertile field for research and development since the 1970s and proven to be efficient in removing irrelevant and redundant features, increasing efficiency in mining tasks, improving mining performance like predictive accuracy, and enhancing comprehensibility of the learned results. The factor selection algorithm performs a search through the space of feature subsets (Allen, 1974). In general, two categories of the algorithm have been proposed to resolve the factor selection problem. The first category is based on a filter approach that is independent of the learning algorithms and serves as a filter to sieve out the irrelevant factors. The second category is based on a wrapper approach, which uses an induction algorithm itself as part of the function evaluating the factor subset (Langley, 1994). Since most of the filter methods are based on a heuristic

Source: Data Mining and Knowledge Discovery in Real Life Applications, Book edited by: Julio Ponce and Adem Karahoca,
ISBN 978-3-902613-53-0, pp. 438, February 2009, I-Tech, Vienna, Austria

algorithm for general characteristics of the data rather than a learning algorithm that evaluates the merits of the factor subsets as done by wrapper methods, filter methods are generally much faster and have more practical capabilities to utilize high dimensionality than wrapper methods.

While a large number of factors are considered, there are three important issues when handling data analysis problems, namely, missing values, outliers, and noise factors. The results from DA that uses a number of data sets including many outliers may often be misleading. An outlier is an observation that lies outside of the overall pattern of a distribution (Bakar et al., 2006). Missing values can often seriously affect the data analysis results if a large number of factors and their associated missing values are ignored. Next, in many scientific and engineering fields, there are a number of data sets that are uncontrollable and difficult to handle, since the nature of the measurement of a performance variable may often be a destructive or very expensive characteristic, which is known as the noise factor (Yang et al., 2007).

Existing studies in DM mostly focus on finding patterns in large data sets and further using them for organizational decision making (Yang et al., 2007). DM methods also may not discuss the robustness of solutions, either by considering data pre-processes for outliers and missing values or by considering uncontrollable noise factors.

In order to address this limitation, we have developed an enhanced DA method incorporating the robust design (RD) principle. Among the process/product design methods currently studied in the science and engineering community, researchers often identify RD as one of the most effective methodologies for process/product improvement. Because of their practicability in reducing the inherent uncertainty associated with input factors and process performance, the widespread applications of RD techniques have resulted in significant improvements in process quality, manufacturability, and reliability at low cost. However, most RD methods reported in the literature may obtain the most favorable solution for a small number of given input control factors without considering the reduction in dimensionality for large databases. Although traditional RD methods consider the selection of potential significant factors when they confront a data set including many factors with an interesting response factor, the process is frequently far from the objective as individual egos because the selection process is based on drawing insight from a number of readily available sources relying on the practitioners' opinion and their experience.

For this reason, we propose an integrated approach called robust data mining (RDM), which can reduce the dimensionality of large data sets, may provide detailed statistical relationships among the factors, and robust factor settings, as shown in Fig. 1. This RDM approach has neither been adequately addressed in the literature nor properly applied in industrial processes. As a result, the primary objective of this paper is three-fold. First, the proposed RDM applies outlier test and expectation maximum (EM) algorithm to carry out the data pre-process. Then, the proposed RDM reduces the dimensionality to find the significant factors among a large number of input factors using correlation-based feature selection (CBFS) method and best first search (BFS) algorithm. These methods can evaluate the worth of a subset including the input factors by considering the individual predictive ability of each factor along with the degree of redundancy between the pairs of input factors. This method is far more effective than any other method when a large number of input factors are considered in a process design procedure. Finally, the proposed model utilizes the theory of robust design to handle noise factors using the concept of surrogate variables and response surface methodology (RSM). Our numerical example clearly shows

that the proposed RDM method can efficiently find significant factors and optimal settings by reducing the dimensionality.

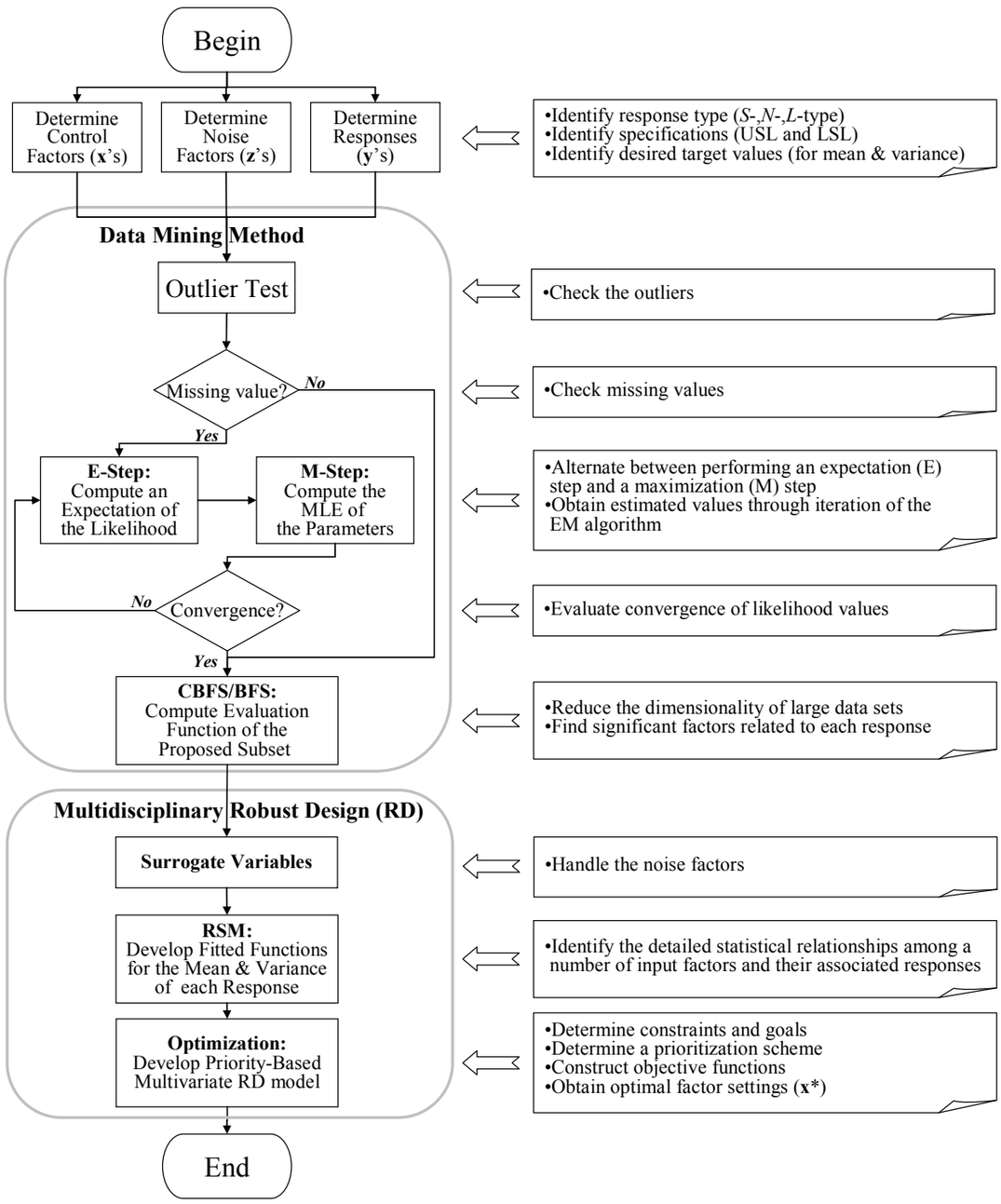

Fig. 1. Overview of the RDM model

2. Stage I: data mining method

2.1 Data pre-process

The issues of outliers and missing values are the two most important problems in the data pre-process procedure. As shown in Fig. 2, the proposed procedure conducts outlier tests to

detect the outliers in a large number of data sets. If the results of the outlier tests include a number of unusual observations, these outliers are deleted and regarded as missing values. To address the missing values, the EM algorithm is utilized.

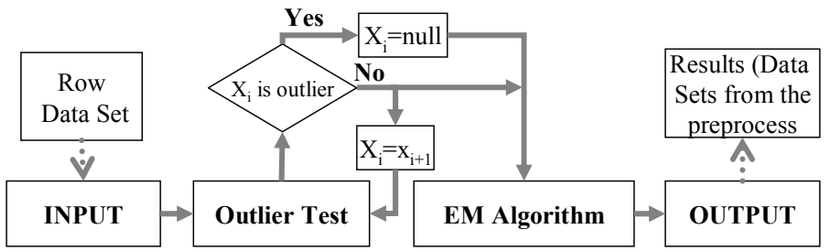

Fig. 2. Proposed data pre-process procedure

2.1.1 Outliers test in data mining

Recently, a number of studies have been conducted on an outlier test for large datasets, which can be categorized into (1) the statistical approach, (2) distance-based approach, and (3) deviation-based approach (Bakar et al., 2006).

The statistical approach to outlier detection assumes a distribution or probability model for the given data set and then identifies the outliers with respect to the model using a discordancy test (Witten & Frank, 2005). One of the drawbacks of the statistical approach is the requirement of knowledge about the parameters of the data set, such as data distribution (Bakar et al., 2006). However, the distance-based approach is based on two parameters that are given in advance using the knowledge about the data or may be changed during the iterations to select the most representative outliers. Deviation-based methods identify the outliers by examining the main characteristics of the objects in a group. Objects that "deviate" from this description are considered outliers. Hence, in this approach, the term deviation is typically used to refer to outliers (Witten & Frank, 2005).

2.1.2 Expectation Maximization (EM) algorithm

The EM algorithm is used in statistics for finding the maximum likelihood estimates of parameters in probabilistic models, where the model depends on the unobserved latent variables (Pernkopf, 2005). The EM alternates between performing an expectation (E) step, which computes the expectation of the likelihood by including the latent variables as if they were being observed, and a maximization (M) step, which computes the maximum likelihood estimates of the parameters by maximizing the expected likelihood found in the E step. The parameters found in the M step are then used to begin another E step, and the process is repeated.

Let Y denote the random vector corresponding to the observed data y, having a probability density function of $g(y; \psi)$, where $\psi = (\psi_1,\ldots, \psi_d)T$ is a vector of unknown parameters within the parameter space Ω. The observed data vector y is viewed as being incomplete and is regarded as an observable function of the complete data. The notion of incomplete data includes the conventional sense of missing data. Let x denote the vector containing the augmented or complete data. Let $g_c(x; \psi)$ denote the probability density function of the random vector X corresponding to the complete-data vector x. Then, the complete-data log-likelihood function that could be formed for ψ if x were fully observable is given by

$$\log L_c(\boldsymbol{\psi}) = \log g_c(\boldsymbol{x};\ \boldsymbol{\psi}) \tag{1}$$

Formally, we have two sample spaces α and β and many-to-one mapping from α to β. Instead of observing the complete-data vector x in α, we observe the incomplete-data vector y = y(x) in β. It follows that

$$g(\boldsymbol{y};\ \boldsymbol{\psi}) = \int_{\alpha(y)} g_c(\boldsymbol{x};\ \boldsymbol{\psi}) d\boldsymbol{x}, \tag{2}$$

where α(y) is the subset of α determined from the equation y = y(x).
The EM algorithm approaches the problem of solving the incomplete-data likelihood function indirectly by proceeding iteratively in terms of the complete-data log likelihood function log Lc(ψ). As this function is unobservable, it is replaced by its conditional expectation given y by using the current fit for ψ. On the (k+1)-th iteration, the E and M steps are defined as follows (McLachlan, 1996):

E-step. Calculate $Q(\boldsymbol{\psi};\ \boldsymbol{\psi}^{(k)})$,
where $Q(\boldsymbol{\psi};\ \boldsymbol{\psi}^{(k)}) = E_{\boldsymbol{\psi}^{(k)}}\{\log\ L_c(\boldsymbol{\psi})|\boldsymbol{y}\}$.
M-step. Choose $\boldsymbol{\psi}^{(k+1)}$ to be any value of $\boldsymbol{\psi} \in \boldsymbol{\Omega}$ that maximizes $Q(\boldsymbol{\psi};\ \boldsymbol{\psi}^{(k)})$; that is,
$Q(\boldsymbol{\psi}^{(k+1)};\ \boldsymbol{\psi}^{(k)}) \geq Q(\boldsymbol{\psi};\ \boldsymbol{\psi}^{(k)})$ for all $\boldsymbol{\psi} \in \boldsymbol{\Omega}$.

The E and M steps are alternated repeatedly until the difference $L(\boldsymbol{\psi}^{(k+1)}) - L(\boldsymbol{\psi}^{(k)})$ changes by an arbitrarily small amount in the case of convergence of the sequence of likelihood values $\{L(\boldsymbol{\psi}^{(k)})\}$. An overview of the EM algorithm is shown in Fig. 3.

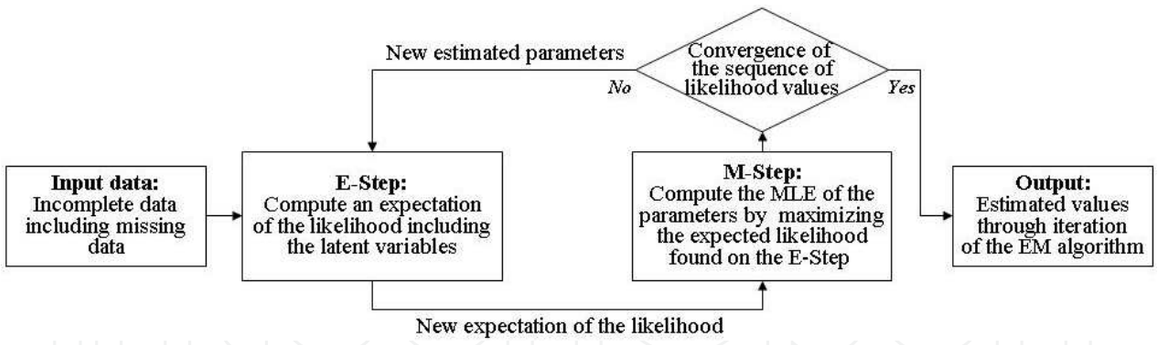

Fig. 3. Overview of the EM algorithm used in data pre-processing

2.2 Data mining procedure

2.2.1 Correlation-Based Feature Selection (CBFS) method

CBFS is a filter algorithm that ranks the subsets of the input features according to a correlation-based heuristic evaluation function. The bias of the evaluation function is toward the subsets that contain a number of input factors, which are not only highly correlated with a specified response but also uncorrelated with each other (Xu et al., 2004). Among the 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 need to 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

$$EV_s = \frac{n\overline{\rho}_{FR}}{\sqrt{n + n(n-1)\overline{\rho}_{FF}}} \tag{3}$$

where EV, $\overline{\rho}_{FR}$, and $\overline{\rho}_{FF}$ represent the heuristic evaluation value of a factor subset S containing n factors, mean of the factor-response correlation (F∈S), and mean of the factor-factor inter-correlation, respectively. Further, $\sqrt{n + n(n-1)\overline{\rho}_{FF}}$ and $n\overline{\rho}_{FR}$ indicate the prediction of the response based on a set of factors and redundancy among the factors, respectively. In order to measure the correlation between two factors or a factor and response, an evaluation of a criterion called symmetrical uncertainty is conducted (Hall, 1998).

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 the factor-factor inter-correlation or factor-response correlation. Unfortunately, information gain is not apt for 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 the bias in information gain toward features with more values and normalize its value within the range [0, 1]. The coefficient of symmetrical uncertainty can be calculated as

$$C_{SU} = 2.0 * \left[\frac{gain}{H(Y) + H(X)} \right] \tag{4}$$

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, probability of y value, conditional entropy of Y given X, and information gain—a symmetrical measure that reflects additional information about Y given X, respectively.

2.2.2 Best First Search (BFS) algorithm

In many literatures, finding the best subset is seldom achieved in many industrial situations when 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 BFS method—a heuristic search method that implements the CBFS algorithm (Langley, 1994). 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 BFS algorithm can backtrack to a more promising previous subset and continue searching from there. The procedure for using the proposed BFS algorithm is given below:

Step 1. Begin with the OPEN list containing the start state, CLOSE list empty, and BEST ← start state (put the start state to BEST).
Step 2. Let a subset θ = arg max EVS (subset), (get the state from OPEN with the highest evaluation EVS).
Step 3. Remove s from OPEN and add to CLOSE.
Step 4. If EVS (θ) ≥ EVS (BEST), then BEST ← θ (put θ to BEST).
Step 5. For each next subset ξ of θ that is not in the OPEN or CLOSE 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.

The evaluation function given in equation (3) is a fundamental element of CBFS that imposes a specific ranking on the factor subsets in the search spaces. In most cases, enumerating all the possible factor subsets is extremely time-consuming. In order to reduce the computational complexity, the BFS method is utilized to find the best subset. The BFS method can start with either no factor or all the 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. The overview of the CBFS and BFS methods is shown in Fig. 4.

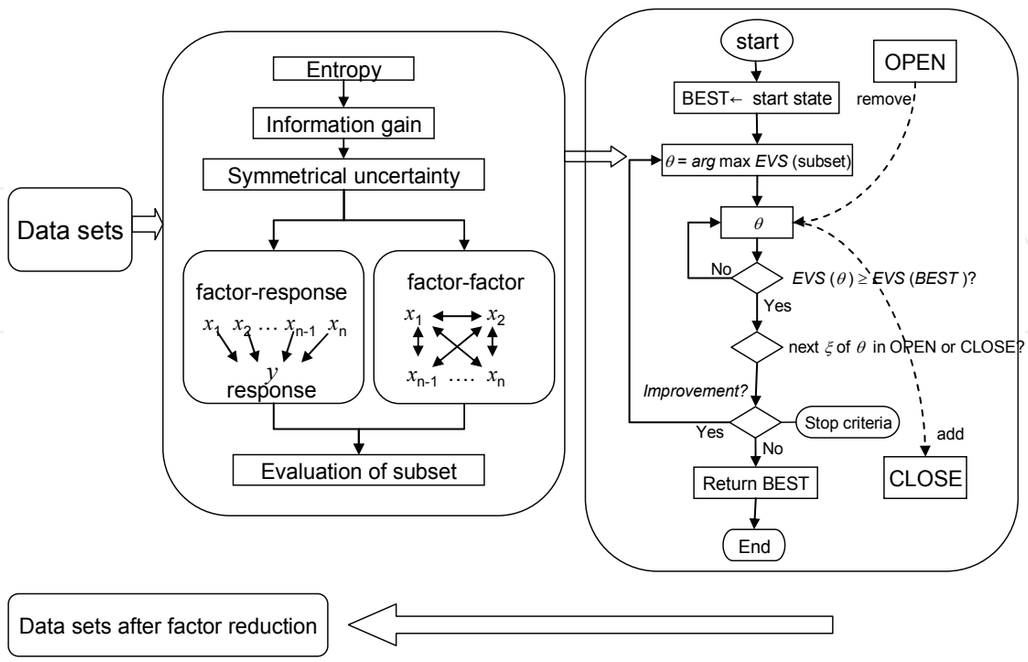

Fig. 4. Overview of the DM method

3. Stage II: robust design

3.1 Surrogate variables

The surrogate variable technique is a subbranch of the screening inspection method. Generally, the nature of measurements or observations on a response (i.e., a dependent variable) may be exceptionally expensive, destructive, or difficult to obtain, forcing a reduction in the overall sample size used to fit the model. To avoid these without dramatically increasing the cost of the experiment, one may use cheaper or more easily collected "surrogate" variables to supplement the expensive input factors.

In our approach, the noise factors of significant factors—both in response and in input—are referred to the destructive or very expensive performance variables to be measured. By using CBFS, we can easily find the candidate surrogate variables from the redundancy factors for every noise factor. Fig. 5 shows two cases of surrogate (k, i, n, m ∈ int). One is when one of the interesting responses y_n exhibits the characteristics of noise, while another interesting response y_m is not only highly correlated to the noise one but also controllable; the surrogate between y_n and y_m can be considered. Another is when we focus on a specific interesting response y_k corresponding to some input factors (x_k, x_i,…, x_n), where factor X_i is noise; however, factor x_1 is neither noise and irrelevant to X_i nor corresponding to the interesting response y_k. Then, x_1 will be the available surrogate variable candidate for x_i.

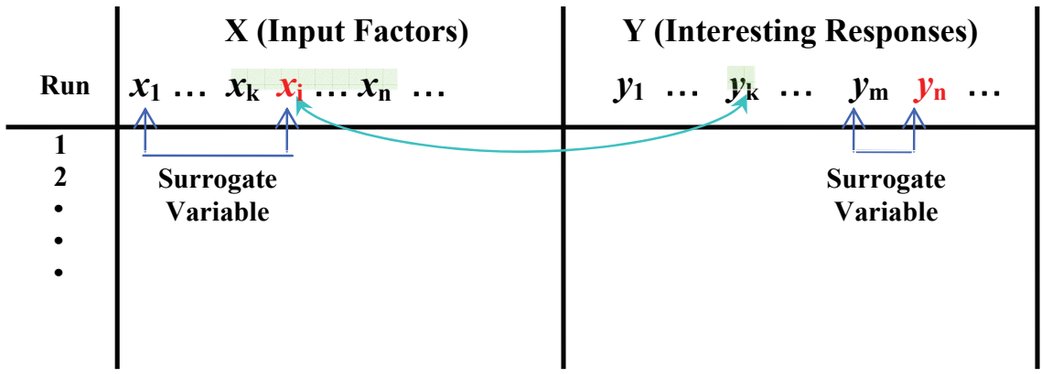

Fig. 5. Concepts of surrogate variables for input factors and responses

3.2 Response Surface Methodology (RSM)

RSM is a statistical tool that is useful for modeling and analyses 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 unknown or is very complicated. For a comprehensive presentation of RSM, Box et al. (Box et al., 1998) and Shin and Cho (Shin & Cho, 2005) provided insightful comments on the current status and future direction of RSM.

In many industrial situations, a manufacturing or service process often contains both control and noise factors that cannot be handled (Montgomery, 2001). Supposing that there are k controllable variables x = $[x_1, x_2, ..., x_k]$ and r noise variables z = $[z_1, z_2, ..., z_r]$, the response model incorporating both control and noise factors can be given by

$$y(\mathbf{x}, \mathbf{z}) = f(\mathbf{x}) + h(\mathbf{x}, \mathbf{z}) + \varepsilon \tag{5}$$

where f(x), h(x, z), and ε denote the portion of the model that involves only the control factors, term involving the main effects of the noise factors and the interactions between the control and noise factors, and random error assumed to be normally distributed with zero mean and certain variance, respectively. The detailed calculation of h(x, z) is

$$h(\mathbf{x},\mathbf{z}) = \sum_{i=1}^{r} \gamma_i z_i + \sum_{i=1}^{k} \sum_{j=1}^{r} \delta_{ij} x_i z_j \tag{6}$$

where γi and δij are the coefficients of noise factors and interactions between the control and noise factors, respectively. Denoting the variance of the noise variables as σ_z^2 and assuming that the noise variables and random errors ε have zero covariance, the mean response model by taking the expectation of the response model in equation (5) can be derived as follows:

$$E_{\mathbf{z}}[y(\mathbf{x},\mathbf{z})] = \hat{\mu}(\mathbf{x}) = f(\mathbf{x}) \tag{7}$$

By using Taylor series expansion, the variance model for the response can be simplified as follows:

$$Var_{\mathbf{z}}[y(\mathbf{x},\mathbf{z})] = \hat{\sigma}(\mathbf{x}) = \sigma_z^2 \sum_{i=1}^{r} \left(\frac{\partial y(\mathbf{x},\mathbf{z})}{\partial z_i} \right)^2 + \sigma^2 \tag{8}$$

where σ^2 is the mean-square error on the analysis of variance (ANOVA).

3.3 Robust Desirability Function (RDF) model

The quality of pharmaceutical products is often judged on multiple responses that are not of the same type. Pharmaceutical quality characteristics typically have one of the three possible goals and are therefore categorized as follows:

1. Smaller-the-better (S type): Minimize the quality characteristic of interest.
2. Nominal-the-better (N type): The quality characteristic of interest has a specific target value.
3. Larger-the-better (L type): Maximize the quality characteristic of interest.

Hence, a special multi-objective optimization model is required. It must be able to handle all the three types of quality characteristics simultaneously and consider robustness to reduce both process bias and variability. To address these issues, we propose a RDF model that can resolve the design problems involving multiple responses of several different types by considering the effect of noise factors. Our proposed model integrates the desire function (DF) that involves a popular approach to formulate and resolve the problem as a multi-objective optimization problem into the mean-squared error (MSE) approach, yielding robust solutions by considering a tradeoff between the process mean and variability. Detailed descriptions on desirability function and MSE model can be found in (Myers, 2002) and (Cho, 1994).

Let S, N, and L represent the indexes of the S-, N-, and L-type quality characteristics, respectively. For MSE_{kS}-related S-type characteristics, the maximum allowable value ($MSE_{kS}^{\max}$) is specified, while for MSE_{kL}-related L-type quality characteristics, the maximum

allowable value ($MSE_{kL}^{\max}$) is specified. It is noted that the maximum value ($MSE_{kN}^{\max}$) needs to be specified for MSE_{kN}-related N-type quality characteristics. Suppose we denote the lower and upper bounds for the control factors as $\breve{x}_i$, $\hat{x}_i$, respectively, and represent the maximum and minimum allowable values for the S-, N-, and L-type quality characteristics as $\breve{y}_{k(S,N,L)}$, $\hat{y}_{k(S,N,L)}$, respectively. Denoting the target values and weights for desirability of the k-th S-, N-, and L-type characteristics by τk(S, N, L) and wk(S, N, L), respectively, we propose the following RDFs:

$$Maximize\ \ D = \left[\prod_{k=1}^{n} d_{kt}\right]^{1/k} \quad for\ t = S-,\ N-and\ L-type \tag{9}$$

where $d_{kt} = \begin{cases} 1 & \text{if } MSE_{kS}(\boldsymbol{x}) \le \tau_{kS},\ MSE_{kN}(\boldsymbol{x}) \le \tau_{kN},\ MSE_{kL}(\boldsymbol{x}) \le \tau_{kL} \\ \left(\dfrac{MSE_{kS}^{\max} - MSE_{kS}(\boldsymbol{x})}{MSE_{kS}^{\max} - \tau_{kS}}\right)^{w_{kS}} & \text{if } \tau_{kS} \le MSE_{kS}(\boldsymbol{x}) \le MSE_{kS}^{\max} \text{ for } k = 1, 2, \ldots, l \\ \left(\dfrac{MSE_{kN}^{\max} - MSE_{kN}(\boldsymbol{x})}{MSE_{kN}^{\max} - \tau_{kN}}\right)^{w_{kN}} & \text{if } \tau_{kN} \le MSE_{kN}(\boldsymbol{x}) \le MSE_{kN}^{\max} \text{ for } k = l+1, \ldots, m \\ \left(\dfrac{MSE_{kL}^{\max} - MSE_{kL}(\boldsymbol{x})}{MSE_{kL}^{\max} - \tau_{kL}}\right)^{w_{kL}} & \text{if } \tau_{kL} \le MSE_{kL}(\boldsymbol{x}) \le MSE_{kL}^{\max} \text{ for } k = m+1, \ldots, n \\ 0 & \text{if } MSE_{kS}(\boldsymbol{x}) \ge MSE^{\max}{}_{kS},\ MSE_{kN}(\boldsymbol{x}) \ge MSE^{\max}{}_{kN},\ MSE_{kL}(\boldsymbol{x}) \ge MSE_{kL}^{\max} \end{cases}$

Constraints $\breve{x}_i \le x_i \le \hat{x}_i$ and $\breve{y}_{kt} \le y_{kt} \le \hat{y}_{kt}$ for $i = 1, 2, \ldots, h$

Note that the objective function D, called the RDF, uses the geometric mean of the individual desirability. It is possible to design a function where the values exceeding the threshold, but still rather less than the target, are only slightly penalized by choosing 0 < w< 1; higher w values are assigned when you need penalize even further. This allows optimization to take into account the relative importance of each quality characteristic or response, while selecting the most appropriate form of the partial desirability function.

4. Numerical example

To effectively demonstrate the implementation of our proposed methodology, actual case studies of processes that produce a placebo tablet have been conducted in which a number of design variables were considered. The data used in this numerical example is obtained from a continuous real-time tablet manufacturing process. The tablet manufacturing process is classified into three stages, namely, flow, compression, and ejection. In the first step,

granules are fed to be compressed into tablets; at the compression stage, granules are compressed into tablets. At the ejection stage, the tablets are ejected. The objective of this study is to commonly optimize each desired bias and variability value of three tablet quality characteristics including friability (y_1), hardness (y_2), and disintegration (y_3). Then, based on prior information about the system under investigation, it logically follows that the first pressure (x_1) to remove air in the granules, second pressure (x_2) to produce tablets, first dwell time (x_3) to remove air in the granules, second dwell time (x_4) to produce tablets, speed to remove the first punch (x_5), speed to remove the second punch (x_6), speed to eject tablets (x_7), amount of overfill (x_8), amount of dust (x_9), and particle size (x_{10}) are the control factors and humidity (z_1) and temperature (z_2) are the noise factors considered in this study. Friability refers to the brittleness of a tablet and it is measured as the percentage of material lost as it passes through a motorized rotary drum. The effect of the motorized rotary drum allows researchers to predict how the tablets will withstand packaging and transportation. Hardness is an important quality characteristic because it is a major concern in tablet manufacturing. A soft tablet will cause problems during compression and a hard tablet can damage teeth. Lastly, disintegration refers to the time (in minutes) that is required for a tablet to dissolve in a suitable liquid at 37°C and is an estimator of how effectively the tablet will release its ingredients within the body. Hardness is measured by applying a uniform force (measured in Newtons) on the tablet until it breaks. In this particular case, the quality characteristics of interest have conflicting objectives, as shown in Table 1. In order to satisfy the goals of all the three quality characteristics, the goal programming approach is used to establish that the hardness objective is the most important and the friability objective is the least important. Table 2 shows the data from the tablet manufacturing process. The data set is an incomplete-data set and each of the five factors—x_1, x_3, x_5, x_7, and x_9—have a missing value.

Quality characteristic	Units	Imp. Rating	Goal	Type	Lower limit	Upper limit
Friability (*F*)	%	3	Minimize	*S*-type	0.4	10
Hardness (*H*)	N	1	Target Value (50)	*N*-type	20	80
Disintegration (*D*)	Min	2	Maximize	*L*-type	0.5	10

Table 1. Quality characteristics of friability (*F*), hardness (*H*), and disintegration (*D*)

No.	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	z_1	z_2	y_1	y_2	y_3
1	4.57	29.69	0.77	2.74	0.27	0.29	0.34	13.23	0.13	134.46	52.08	20.82	4.14	38.79	6.54
2	4.48	38.51	0.75	3.82	0.26	0.47	0.48	21.33	0.12	139.32	41.28	27.42	7.92	67.24	13.19
3		30.58	0.65	4.51	0.23	0.31	0.56	14.04	0.13	138.78	57.60	21.12	4.68	40.71	8.23
4	4.66	32.34	0.78	4.47	0.27	0.53	0.56	23.76		140.67	49.44	22.14	6.66	63.55	11.28
5	3.77	34.99	0.63	3.90	0.22	0.28	0.49	12.42	0.15	138.24	52.32	24.06	8.55	82.50	14.17
6	3.95	29.84		4.11	0.23	0.46	0.51	20.79	0.13	135.54	61.68	16.74	3.72	33.62	6.74
7	4.48	42.19	0.75	3.63	0.26	0.51	0.45	22.95	0.14	132.30	63.36	23.52	8.58	82.48	14.17
8	4.40	27.64	0.74	4.60	0.26	0.31		14.04	0.12	138.78	53.28	21.48	6.54	57.39	11.07
9	5.02	28.37	0.84	3.59		0.41	0.45	18.36	0.13	131.76	51.60	21.42	4.95	46.52	8.59
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
100	5.02	32.78	0.84	3.02	0.30	0.32	0.38	14.31	0.15	148.23	57.36	22.98	8.12	47.03	4.56

Table 2. Data set for the tablet manufacturing process

4.1 Stage I: EM algorithm for data pre-process

DA using the EM algorithm provides the results of the data pre-treatment for the missing values in order to conduct the DM procedure. Table 3 lists the estimated mean, standard deviation, and value after performing 25 iterations of the EM algorithm using the SPSS software package. Consequently, reasonable values based on the estimated mean, standard deviation, and covariance parameters for the five missing values among the twelve factors are found to be (x_1, x_3, x_5, x_7, x_9) = (3.86, 0.66, 0.30, 0.58, 0.14).

Significant factors	Estimated mean	Estimated standard deviation	Estimated values
x_1	4.44	0.35	3.86
x_3	0.74	0.06	0.66
x_5	0.26	0.02	0.30
x_7	0.49	0.10	0.58
x_9	0.13	0.01	0.14

Table 3. Estimated mean, standard deviation, and value after performing 25 iterations of the EM algorithm

4.2 Stage II: DM for dimensionality reduction

The CBFS method, a DM technique, was used to seek the highly correlated factors associated with interesting responses (i.e., friability, hardness, and disintegration) by reducing the dimensionality related to a large number of factors and removing irrelevant and redundant data. As shown in Table 4, data mining results obtained using the Weka software package indicate that two uncorrelated factors (i.e., x_2 and z_2), four uncorrelated factors (i.e., x_1, x_2, x_9, and z_2), and three uncorrelated factors (i.e., x_2, x_9, and z_2) are significant for y_1, y_2, and y_3, respectively. Among these solutions, the temperature (z_2) often cannot be controlled in the tablet manufacturing process. Consequently, we consider z_2 as the noise factor, and consider the others input factors (i.e., x_1, x_2, and x_9) among the DM results as the control factors.

DM	Responses		
	y_1	y_2	y_3
Search method	Best first	Best first	Best first
Search direction	forward	forward	forward
Total number of subsets evaluated	64	79	67
Selected factors	x_2, z_2	x_1, x_2, x_9, z_2	x_2, x_9, z_2

Table 4. DM results for responses y_1, y_2, and y_3

4.3 Stage III: results of multidisciplinary RD using RSM

Based on the results of the significant factor selection, RSM was performed by using the MINITAB software package to identify comprehensive relationships among a large number of factors and their associated responses. DA using the RSM provides the following fitted polynomial models for each quality characteristic:

$$\begin{aligned} y_{1S} = & -18.63+14.84x_1+0.07x_2-175.15x_9-0.30z_2-0.77x_1^2+0.02x_2^2+2840.08x_9^2 \\ & +0.34x_1x_2-139.15x_1x_9-0.01x_1z_2-11.37x_2x_9-0.05x_2z_2+18.24x_9z_2 \end{aligned} \quad (10)$$

$$y_{2N} = -149.00+132.40x_1 -1.50x_2 -1665.40x_9 -3.20z_2 -6.70x_1^2 +0.20x_2^2 +26047.70x_9^2 \\ +3.20x_1x_2 -1281.40x_1x_9 -0.00x_1z_2 -90.90x_2x_9 -0.50x_2z_2 +167.30x_9z_2 \quad (11)$$

$$y_{3L} = -22.35+21.98x_1 +0.07x_2 -297.09x_9 -0.56z_2 -1.16x_1^2 +0.03x_2^2 +4477.10x_9^2 \\ +0.55x_1x_2 -216.36x_1x_9 +0.00x_1z_2 -17.50x_2x_9 -0.08x_2z_2 +28.53x_9z_2 \quad (12)$$

The response models for y_{1S}, y_{2N}, and y_{3L} are adequate for use as a response function since the results yield 76.2, 77.4, and 76.1% R-sq, respectively. Let $\sigma_{z_2}^2$, σ_{1S}^2, σ_{2N}^2, and σ_{3L}^2 denote the variance of the noise variables and mean-square error of the ANOVA values for friability, hardness, and disintegration, respectively. Using equations (10)–(12), the fitted polynomial models of the mean and variance can be written as

$$\hat{\mu}_{1S}(\mathbf{x}) = -18.63+14.84x_1 +0.07x_2 175.15x_9 -0.77x_1^2 +0.02x_2^2 +2840.08x_9^2 \\ +0.34x_1x_2 -139.15x_1x_9 -11.37x_2x_9 \quad (13)$$

$$\hat{\sigma}_{1S}(\mathbf{x}) = 1.23+0.04x_1 +0.19x_2 -69.39x_9 +0.63\text{e-}3x_1^2 +0.63\text{e-}2x_1x_2 -2.31x_1x_9 \\ +0.02x_2^2 -11.56x_2x_9 +2109.30x_9^2 \quad (14)$$

$$\hat{\mu}_{2N}(\mathbf{x}) = -149.00+132.40x_1 -1.50x_2 -1665.40x_9 -6.70x_1^2 +0.20x_2^2 \\ +26047.70x_9^2 +3.20x_1x_2 -1281.40x_1x_9 -90.90x_2x_9 \quad (15)$$

$$\hat{\sigma}_{2N}(\mathbf{x}) = 122.10+20.29x_2 -6788.37x_9 +1.59x_2^2 -1060.68x_2x_9 +177452.10x_9^2 \quad (16)$$

$$\hat{\mu}_{3L}(\mathbf{x}) = -22.35+21.98x_1 +0.07x_2 -297.09x_9 -1.16x_1^2 +0.03x_2^2 +4477.10x_9^2 \\ +0.55x_1x_2 -216.36x_1x_9 -17.50x_2x_9 \quad (17)$$

$$\hat{\sigma}_{3L}(\mathbf{x}) = 3.56+0.57x_2 -202.59x_9 +0.04x_2^2 -28.94x_2x_9 +5160.51x_9^2 \quad (18)$$

where $\hat{\mu}_{1S}(\mathbf{x})$, $\hat{\sigma}_{1S}(\mathbf{x})$, $\hat{\mu}_{2N}(\mathbf{x})$, $\hat{\sigma}_{2N}(\mathbf{x})$, $\hat{\mu}_{3L}(\mathbf{x})$, and $\hat{\sigma}_{3L}(\mathbf{x})$ represent the fitted polynomial models for the mean and variance of friability, hardness, and disintegration, respectively. $\sigma_{z_2}^2$, σ_{1S}^2, σ_{2N}^2, and σ_{3L}^2 are 6.34, 0.66, 57.18, and 1.57 from the ANOVA result for each quality characteristic, respectively. Equations (13)–(18) are used in the proposed RDF model. The target value for the process mean of friability, hardness, and disintegration are 0.4, 50, and 10, respectively (i.e., τ_{1S} = 0.4, τ_{2N} = 50, and τ_{3L} = 10]. Additionally, the constraints on x_1, x_2, and x_9 can be expressed as

$$3 \le x_1 \le 6, \quad 24 \le x_2 \le 45, \quad 0.1 \le x_9 \le 0.2 \quad (19)$$

We then convert the three quality characteristics into three MSEs of the same type. The target values and upper limits for the three MSE models are 0, 0, 0, 403.34, 3101.21, and

143.01, respectively. Suppose that the weights for the desirability of MSEs based on the friability, hardness, and disintegration are 1 (i.e., $w_{(1S,2N,\text{ and }3L)}$ = 1).

By utilizing the proposed RDF model, which makes the multi-objective optimization problem inherently easier to solve due to the fact that single-objective optimization approaches can be applied, the multi-objective optimization problem can be transformed into a single response problem. In order to resolve this single response problem to maximize the geometric mean of the individual desirability of MSEs, MINITAB can be used. Using the MINITAB package, the optimal solutions are found to be (x_1^*, x_2^*, x_9^*) = (4.847, 41.315, 0.160). The optimal solution and predicted value of the mean and variability of each process for this case are listed in Table 10.

Ingredients	Optimal Solution	Quality Characteristics	Predicted Value	
			$\hat{\mu}(\mathbf{x})$	$\hat{\sigma}^2(\mathbf{x})$
First pressure (x_1)	4.847	Friability (y_1)	1.934	2.272
Second pressure (x_2)	41.315	Hardness (y_2)	61.280	110.890
Amount of dust (x_9)	0.160	Disintegration (y_3)	4.780	4.673

Table 10. Optimal solutions for the tablet manufacturing process

7. Conclusion

In this paper, we developed a RDM method by integrating a DM method for pre-processing unclear data and finding significant factors into a multidisciplinary RD method for providing the best factor settings. Based on the results of the DM method, we found important factors for placebo tablet manufacturing among a large data set. By using the BFS method, the CFBS method in its pure form is exhaustive, but the use of a stopping criterion expedites the probability of searching the entire data set. We then conducted RD optimization using the RSM and RDF methods, while incorporating an uncontrollable noise factor. We finally showed that the proposed RDM method could efficiently find significant factors and optimal settings by reducing the dimensionality through the numerical example. In order to examine the proposed RDM method, the consideration of different case studies can be a possible future research issue.

8. Acknowledgement

This work was supported by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korea government (MOST) (No. R01-2007-000-21070-0).

9. References

Allen, D. (1974). The Relationship between Variable Selection and Data Augmentation and a Method for Prediction, *Technometrics*, Vol. 16, No. 1, (Feb. 1974) 125–127

Bakar, Z.A.; Mohemad, R.; Ahmad, A. & Deris, M.M. (2006). A Comparative Study for Outlier Detection Techniques in Data Mining, *proceedings of IEEE Conference on Cybernetics and Intelligent Systems,* pp. 1-6, ISBN: 1-4244-0023-6, Bangkok, Thailand, June 2006, IEEE Press New York

Box, G.E.P.; Bisgaard, S. & Fung, C. (1998). An Explanation and Critique of Taguchi's Contributions to Quality Engineering. *International Journal of Reliability Management*, Vol. 4, No. 2, (Jan. 1998) 123-131, ISSN: 1099-1638

Cho, B.R. (1994). Optimization issues in quality engineering, Ph.D. dissertation, School of Industrial Engineering, University of Oklahoma, OK, U.S.

Gardner, M. & Bieker, J. (2000). Data Mining Solves Tough Semiconductor Manufacturing Problems. *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 376-383, ISBN: 1-58113-233-6, Boston, U.S., Aug. 2000, ACM, New York

Hall, M. A. (1998). Correlation-based Feature Selection for Machine Learning. Ph.D Dissertation, Waikato University, Department of Computer Science. Hamilton, New Zealand

Langley, P. (1994). Selection of Relevant Features in Machine Learning, *Proceedings of the AAAI Fall Symposium on Relevance*, pp. 140-144, ISBN 978-0-929280-76-9, New Orleans, U.S., Nov. 1994, AAAI Press, U.S.

McLachlan, G. J., & Krishnan, T. (1996) The EM algorithm and extensions, John Wiley & Sons, New York, ISBN: 978-0-471-12358-3, New York

Montgomery D.C. (2001). *Introduction to Statistical Quality Control*, John Wiley & Sons, ISBN: 0-471-39412-2, New York

Myers, R. H. & Montgomery, D. C. (2002). *Response surface methodology: process and product optimization using designed experiments*, John Wiley & Sons, ISBN: 978-0-470-17446-3, New York

Pernkopf, F.; Bouchaffra, D. (2005). Genetic-based EM algorithm for learning Gaussian mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 27, No. 8, (Aug. 2005) 1344 – 1348, ISSN:0162-8828

Su, C.T., Chen, M.C., & Chan, H.L. (2005). Applying Neural Network and Scatter Search to Optimize Parameter Design with Dynamic Characteristics. *Journal of the Operational Research Society,* Vol. 56, No. 10, 1132-1140, ISSN 0160-5682

Shin, S. & Cho, B.R. (2005). Bias-specified robust design optimization and its analytical solutions. Computer & Industrial Engineering, Vol. 48, No. 1, (Jan. 2005) 129-140, ISSN: 0360-8352

Shin, S.; Guo Y.; Choi Y. & Choi M. (2006). Development of a Robust Data Mining Method Using CBFS and RSM, *LNCS 4378*, pp. 337-388, ISBN: 978-3-540-70880-3, Novosibirsk, Russia, June 2006, Spriger-Verlag, Berlin Heidelerg

Witten, I.W.H. & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, ISBN: 0-12-088407-0, San Francisco

Xu, Q.; Kamel, M. & Salama, M.M.A. (2004). Significance Test for Feature Subset Selection on Image Recognition, *LNCS 3211*, pp.244-252, ISBN: 978-3-540-23223-0, Porto, Portugal, Sept. 2004, Spriger-Verlag, Berlin Heidelerg

Yang, L.; Shin, S.; Choi, Y.; Choi, M. & Lee, Y. (2007). A Surrogate Variable-Based Data Mining Method using CFS and RSM, *Proceedings of the 6th WSEAS International Conference on Applied Computer Science*, pp.651-657, ISBN: 978-960-8457-61-4 , Hangzhou, China, April 2007, World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, U.S.

Yang, L.; Shin, S. ; Choi, Y. ; Park, K.; Kaewkuekool, S.; Chantrasa, R. & Lila, B. (2007). Development of an Extended Robust Data Mining (ERDM) Model, *Proceedings of*

International Conference on Control, Automation and Systems, pp. 1523-1528, ISBN: 978-89-950038-6-2, Seoul, Korea, Oct. 2007, IEEE Press, New York

Yu, L. & Liu, H. (2003). Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution, *Proceedings of the 20th International Conference on Machine Leaning*, pp. 856-863, ISBN 978-1-57735-189-4, Washington D.C., U.S., Aug. 2003, AAAI Press, U.S.

Data Mining and Knowledge Discovery in Real Life Applications

Edited by Julio Ponce and Adem Karahoca

ISBN 978-3-902613-53-0
Hard cover, 436 pages
Publisher I-Tech Education and Publishing
Published online 01, January, 2009
Published in print edition January, 2009

This book presents four different ways of theoretical and practical advances and applications of data mining in different promising areas like Industrialist, Biological, and Social. Twenty six chapters cover different special topics with proposed novel ideas. Each chapter gives an overview of the subjects and some of the chapters have cases with offered data mining solutions. We hope that this book will be a useful aid in showing a right way for the students, researchers and practitioners in their studies.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Sangmun Shin, Le Yang, Kyungjin Park and Yongsun Choi (2009). Robust Data Mining: An Integrated Approach, Data Mining and Knowledge Discovery in Real Life Applications, Julio Ponce and Adem Karahoca (Ed.), ISBN: 978-3-902613-53-0, InTech, Available from: http://www.intechopen.com/books/data_mining_and_knowledge_discovery_in_real_life_applications/robust_data_mining__an_integrated_approach

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen