Journal of the Society of Korea Industrial and Systems Engineering
Vol, 30, No. 2, pp.99－105, June 2007.

데이터 마이닝을 이용한 로버스트 설계 모형의 최적화

정혜진* · 구본철**†

*동아대학교 산업경영공학과
**동명대학교 기계공학과

Optimization of Robust Design Model using Data Mining

Hey-Jin Jung* · Bon-Cheol Koo**†

*Department of Industrial & Management Systems Engineering, Dong-A University
**Department of Mechanical Engineering, Tongmyong University

According to the automated manufacturing processes followed by the development of computer manufacturing technologies, products or quality characteristics produced on the processes have measured and recorded automatically. Much amount of data daily produced on the processes may not be efficiently analyzed by current statistical methodologies (i.e., statistical quality control and statistical process control methodologies) because of the dimensionality associated with many input and response variables. Although a number of statistical methods to handle this situation, there is room for improvement. In order to overcome this limitation, we integrated data mining and robust design approach in this research. We find efficiently the significant input variables that connected with the interesting response variables by using the data mining technique. And we find the optimum operating condition of process by using RSM and robust design approach.

Keywords : Data Mining, CBFS, BFS, RSM, Robust Design

1. 서 론

오늘날 제조 공정 시스템이 자동화됨에 따라 하루에도 수천 수억 개의 품질 특성치들이 계측된다. 이렇게 발생한 데이터들은 데이터베이스화 되어 실시간으로 공정의 상태를 파악하는데 사용되어진다. 이 데이터로부터 유용한 정보를 빨리 찾고 분석하여 공정 설계에 반영함으로써, 공정의 문제점을 찾아내어 공정을 개선시킬 수 있다. 과거부터 지금까지 제조업에서는 공정과 제품의 품질을 개선하기 위하여 통계적 공정관리(SPC)와 통계적 품질관리(SQC) 기법들을 사용하였다. 그리고 기술자들에 의해 지정된 인자들에 대해 실험계획이나 샘플링 기법들을 실시하여 데이터를 얻어서 공정을 관리하였다. 하지만 이러한 제조 환경에서 기존의 통계적 기법을 사용하여 공정을 관리하는 것은 현실적으로 많은 어려움이 발생한다. 왜냐하면 통계적 기법은 적은 양의 데이터를 정확하게 분석하지만, 인자들이 많아지고 데이터 수가 방대해지면 분석이 어려워지기 때문이다[3]. 그러나 자동으로 계측된 데이터들은 데이터베이스화 되어 수많은 인자들과 데이터들을 포함하고 있다. 따라서 본 연구에서는 이러한 문제점을 해결하기 위한 하나의 대안으로 데이터 마이닝 기법을 사용하였다.

데이터 마이닝 기법이 통계학 분야에서 각광받지 못하는 이유는 찾아낸 패턴들을 임의적인 현상일 수 있다는 불확실성과 최적해를 제시하지 못한다는 이유 때문이다[2]. 본 연구에서는 로버스트 설계와 결합시켜 최적해를 제공함으로써 데이터 마이닝의 단점을 보완해 줄 것이다. 본 연구에서는 이러한 통계적 기법들과 데이터

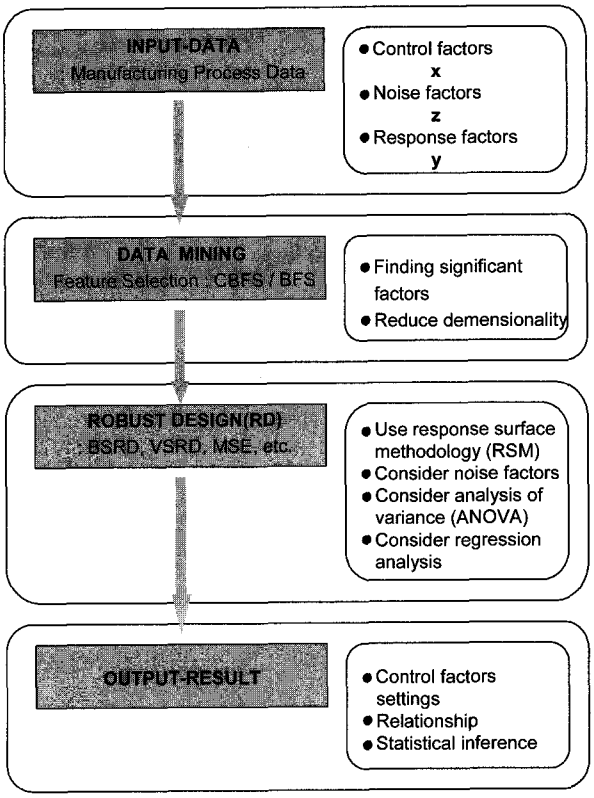
† 교신저자 bckoo@tu.ac.kr

마이닝의 단점들은 보완하면서 장점들은 절충하는 보다 실용적인 새로운 통계적 공정 설계 기법을 제시하고자 한다.

기존의 데이터 마이닝에 대한 연구들은 크게 두 가지로 나눌 수 있는데, 하나는 적용 사례에 관한 연구이고 다른 하나는 데이터 마이닝 알고리즘 개발 및 비교 평가 연구이다. 첫 번째 경우에는, 보유 고객의 과거 자료를 기반으로 여러 가지 데이터 마이닝 기법에 의해 모형을 구축하고 이 모형을 바탕으로 금융기관의 개인 신용평가나 신용카드 회사의 사기감지, 통신 서비스회사의 고객 이탈 방지 등을 예측한 사례들이 제시되었다. 대부분 사회 과학 분야에서 각광받고 있던 데이터 마이닝 기법들이 최근에는 제조 공정에서도 많이 적용되어지고 있다. Gardner and Bieker[9]는 반도체 공정에 데이터 마이닝 기법을 적용하였고, Chang and Xian[5]은 널링 공정의 예측 모형에 데이터 마이닝 기법을 적용하였다. 이현우 등[2]이 LCD 산업에서의 제조공정에 데이터 마이닝 기법을 적용하여 발생할 수 있는 문제점과 해결방안을 제시하고 있다.

두 번째 경우에는 기존의 알고리즘을 비교 평가하거나 새로운 알고리즘을 개발하여 제시하는 연구들이다. 이러한 알고리즘의 개발과 분석에 관한 대표적인 사례로는 Cho et al.[6]은 다수의 의사결정 기준을 갖는 절차에서 연관규칙의 우선권에 관해 연구하였다. DuMouchel [8]는 대형 도수분포표에서 베이지안 데이터 마이닝 기법을 개발하여 이를 FDA(Food and Drug Administration) 자발적 보고 시스템에서 질병과 약의 관계 파악에 응용하였다. 김만선, 이상용[1]에 의해 신경망을 이용한 대용량 데이터 처리를 위한 군집화 기법에 관한 연구가 행해졌다. 그러나 기존의 데이터 마이닝 기법의 적용에 관한 연구들은 여러 알고리즘을 사용하여 인자들 간의 관련성을 규명하는데 만 초점을 두었다. 본 연구에서는 인자들 간의 관련성을 규명하는데 그치는 것이 아니라 한 단계 발전하여 관련성이 높은 인자들을 통계적 기법들과 결합시켜 더 많은 정보를 얻고자 한다.

로버스트 설계에 대한 개념은 다구찌에 의해 처음으로 제시되었다. 통계적 기법과 통합된 다구찌 로버스트 이론은 많이 연구되었다[4, 14]. Vining and Myers[8]는 실험적 데이터를 기초로 한 공정 평균과 분산이 분리된 모형인 쌍체 반응(dual response) 모형을 개발하였다. Cho, Lin and Tu[12]는 평균 제곱 오차(MSE : mean squared error) 모형으로 불리는 모형을 개발하였다. MSE 모형은 모든 조건에서 동일한 우선권을 부여하기 때문에 공정 편의(bias)와 변동 사이의 균형(trade-off)에 관하여 유연성을 제공하지 못하기 때문에 Cho et al.[7]는 이를 해결하기 위해 공정 편의와 변동에 가중치를 부여한 모형을 제시하였다.



<그림 1> 데이터 마이닝을 이용한 로버스트 설계 모형의 최적화 과정

본 연구에서는 앞에서 제시한 데이터 마이닝과 통계적 기법들에 대한 문제점을 해결하는 대안을 제시하기 위하여 데이터 마이닝을 이용한 로버스트 설계 모형을 개발하였다. <그림 1>에서 보여주는 것처럼 크게 두 단계로 나눌 수 있다.

첫 번째 단계는 제조공정 데이터들의 인자 x, z, y를 정의하고, 데이터 마이닝 기법 중 특성선택 알고리즘 CBFS(Correlation-based Feature Selection)를 이용하여 대량의 데이터와 인자들을 포함하는 공정 데이터로부터 반응치 y에 영향을 미치는 주요 인자들을 선택한다.

두 번째 단계에서는 반응표면분석(RSM : Response Surface Methodology)을 실시하여 반응치 y와 선택된 주요 인자들 간의 함수 관계를 추정한 후, 공정의 평균과 분산을 구해서 로버스트 설계(RD : Robust Design) 모형을 세운다. 그리고 로버스트 모형에 대하여 잡음인자에 둔감한 공정의 최적가동조건을 구한다. 첫 번째 단계에서는 변화하는 제조 공정에 데이터 마이닝 기법을 사용함으로써 인자 선택과 비용의 효율성을 제공해 주며 두 번째 단계에서는 데이터 마이닝이 갖는 패턴의 불확실성과 최적해에 관한 문제점을 해결 해 준다.

2. 데이터 마이닝을 이용한 로버스트 설계의 최적화

2.1 데이터 마이닝과 CBFS 알고리즘

여러 학자들에 의해 데이터 마이닝은 다양하게 정의되어지고 있다. "데이터 마이닝은 대용량의 데이터로부터 이들 내에 존재하는 관계, 패턴, 규칙 등을 탐색하고 찾아내어 모형화 함으로써 유용한 지식을 추출하는 일련의 과정(process)이다"[15, 17]. 마케팅, 유통업, 은행업, 보험업 등과 같이 사회경영과학 분야에서 각광받고 있다. 그러나 최근에는 품질 특성치들을 관리하고 분석하기 위하여 제조업 분야에서도 많이 사용되어지고 있다. 실제로 데이터 마이닝 기법을 통해 결함률이 3~5%이었던 업체가 0.03~0.05%로 줄여진 연구 사례도 있다[19].

과거에는 기술자나 전문가에 의해 지정된 인자들의 데이터를 샘플링하거나 실험계획을 실시하여 공정의 불량을 파악하고 원인을 찾고자 하였다. 그러나 최근에는 제조 공정 환경의 자동화로 인하여 대부분의 제품에 대하여 많은 품질 특성치들이 자동으로 측정되어지고 있다. 데이터 마이닝 기법을 통해 이 거대한 현장 데이터로부터 반응치에 영향을 미치는 주요 인자를 알아내고, 이 인자들을 분석하여 공정을 관리한다면 보다 더 정확한 결과를 얻을 수 있으며, 우리가 과거 자료로부터 인식하지 못하고 있던 잠재원인들도 찾을 수 있을 것이다. 뿐만 아니라 별도의 샘플링이나 실험의 비용을 들이지 않고서도 데이터를 얻을 수 있기 때문에 비용적인 면에서도 효율적이다.

데이터 마이닝은 거대한 데이터베이스로부터 반응치(response)에 유의하게 영향을 미치는 인자들을 선택하여 공정의 변화를 설명할 수 있다. 데이터 마이닝은 데이터베이스(database), 기계 학습(ML : machine learning), 지식정보시스템(intelligent information system), 통계학과 전문가 시스템과 같은 분야와 관련되어 있다. 기계는 수억 수천 개의 데이터로부터 정보를 얻기 위해서는 많은 시간과 능력이 요구된다. 거대한 데이터로부터 상관관계가 없거나 중복된 데이터를 제거하여 데이터의 차원을 줄임으로써 기계학습 알고리즘의 수행도를 개선시킬 수 있다. 이를 위해 특성 선택(feature selection)을 기계학습 알고리즘을 적용하기 전에 데이터에 적용한다.

최근 연구에 의해 특성 선택(feature selection)이 기계학습 알고리즘의 수행도에 대하여 긍정적인 영향을 미친다고 증명되었다[10]. 어떤 알고리즘은 상관이 없거나 중복된 데이터를 가지는 많은 양의 데이터들에 의해 수행도가 너무 늦어지거나 또는 불리하게 이행되어지는 경우가 있다. 이런 경우에 특성 선택은 기계학습 알고리즘의 수행도를 향상시켜주고, 가설 탐색 공간을 줄여주며, 저장 필요조건을 줄여준다. 특성선택 알고리즘 기법은 filter와 wrapper 두 가지가 있다. filter는 상관없는 인자와 중복된 인자를 걸러주는 여과기와 같은 방법으로 인자를 선택하고, wrapper 기법은 인자 특성들을 평가하기 위하여 크로스 확인(cross-validation)과 같은 통계적 재샘플링 기법에 따라 인자를 선택해 주는 기법이다. 주요한 인자 집합을 평가하는 학습 알고리즘을 사용하는 wrapper 기법보다 데이터의 일반적인 특성을 기초로 한 휴리스틱 방법을 사용하는 filter 기법이 보다 빠르며 높은 차원의 데이터에 대하여 보다 실용적이기 때문에 더 선호 된다[11]. 따라서 본 연구에서도 filter 기법 중 하나인 CBFS 사용하여 주요 특성을 선택하고자 한다.

CBFS(Correlation-based Feature Selection)는 휴리스틱 평가함수를 기초로 한 상관관계에 따라 입력 특성의 부분 집합에 대하여 순위를 매기는 filter 알고리즘이다. 평가 함수의 기본 개념은 지정된 반응치에 높은 상관관계를 갖는 특성뿐만 아니라 서로 상관관계를 갖지 않는 모든 인자들을 포함하는 부분집합에 대해서 적용된다. 입력 특성들 중에서 관계가 없는 인자들은 무시하고, 비록 반응치에 높은 상관관계가 있다할지라도 중복된 인자들은 제거한다. 특성의 선택은 예측에 사용되지 않은 사례의 영역에서 반응치를 예측하는 정도에 의존한다. 제시된 부분집합의 평가함수는 식 (1)과 같다.

$$EV_s = \frac{n\bar{r}_{FR}}{\sqrt{n+n(n-1)\bar{r}_{FF}}} \quad \cdots\cdots (1)$$

여기서 EV_s는 n개의 인자를 포함하고 있는 특성의 부분 집합 S의 휴리스틱 평가 값을 나타내고, $\bar{r}_{FR}$은 특성과 반응치의 상관관계의 평균치를 나타내며, $\bar{r}_{FF}$는 특성과 특성의 교호상관관계의 평균치를 나타낸다. $\sqrt{n+n(n-1)\bar{r}_{FF}}$와 $n\bar{r}_{FR}$는 각각 특성 사이의 중복성과 특성의 집합을 기초로 한 반응치의 예측을 나타낸 것이다. 두 개의 특성들 사이의 상관관계를 측정하기 위해서 대칭의 불확실성(symmetrical uncertainty)이라 불리는 평가 기준을 구한다.

대칭성 측정은 X가 관측된 후 Y에 대해 얻어지는 정보량과 Y가 관측된 후 X에 대해 얻어지는 정보량이 같음을 나타낸 것이다. 대칭성은 특성과 특성의 교호상관관계 또는 특성와 반응치의 상관관계의 측정을 위해 요구되는 특성이다. 또한 $\bar{r}_{FR}$와 $\bar{r}_{FF}$는 비교 가능하고 동일한 효과를 가지도록 표준화시켜야 한다. 대칭의 불확실성은 보다 가치 있는 특성에 대해 얻어진 정보의 편

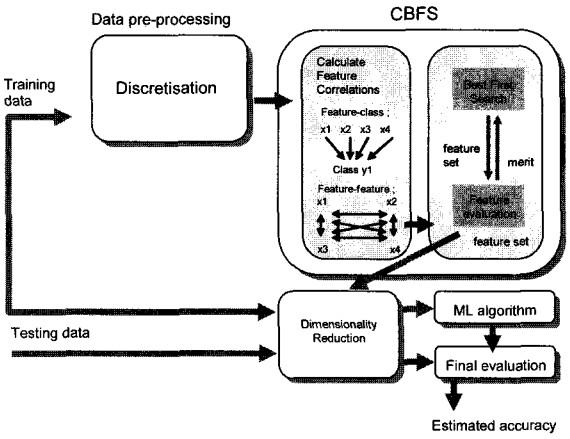
의를 최소화하기 위해 범위 [0,1]에서 표준화한다. 대칭의 불확실성의 상관계수는 식 (2)와 같이 계산되어진다.

$$r_{FF} = \text{대칭의 불확실성}$$
$$= 2.0 \times \left[\frac{gain}{H(Y)+H(X)} \right] \quad \cdots\cdots (2)$$

여기서

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

여기서 $H(Y)$는 지정된 반응치 Y의 엔트로피(entropy)를 나타내고, $p(y)$는 y의 확률을 나타내며, $H(Y|X)$는 X가 주어졌을 때 Y의 조건부 엔트로피이고, $gain$은 대칭성인 X가 주어졌을 때 Y에 대해 추가된 정보량을 얼마나 반영하는지를 측정하여 얻은 정보이다.



<그림 2> CBFS와 BFS 알고리즘

모두를 열거하는 방법을 사용하여 가장 좋은 부분집합을 구하는 것은 거의 불가능하다. 수많은 부분집합을 평가하기 위해 탐색공간을 줄이는 가장 효율적인 방법 중에 하나가 BFS(Best First Search) 기법이다[20]. 이 탐색 방법은 위에 있는 CBFS 알고리즘을 수행하기 위한 휴리스틱 탐색기법이다. 이는 탐색 경로를 따라 뒤로 돌아가는(backtracking)것을 허용하는 개선된 탐색 전략이다. 만약 탐색하던 경로가 가망성이 부족해 보이면, BFS는 가망성이 보다 높아 보이는 이전의 부분집합으로 되돌아간다. 식 (1)에서 주어진 평가함수는 탐색공간에서 특성 부분집합에 대하여 특정한 순위를 부과하기 위한 CBFS의 기본 요소이다. 모든 가능한 인자를 다 열거하는 것은 천문학적인 시간을 소요한다. 계산적인 복잡성을 줄이기 위하여 BFS 기법은 가장 좋은 집합을 찾고자 할 때 유용하게 사용되어진다.

전자 탐색(former search)은 탐색 공간을 통해 결과에 하나의 인자를 추가해 가면서 앞쪽으로 나아가는 방법이고 후자 탐색(latter search)은 탐색 공간을 통해 결과로부터 하나의 인자를 빼가면서 뒤로 가는 방법이다. 탐색 공간을 전체 탐색하는 것을 막기 위해서는 끝내는 기준을 부과해야 한다. 충분히 전개된 부분집합 5개가 연속적으로 현재 가장 좋은 부분집합에 대하여 개선을 보이지 않으면 탐색과정은 끝을 낸다. <그림 2>은 위에서 설명한 CBFS와 BFS의 알고리즘 절차를 나타낸 것이다.

2.2 반응표면분석(RSM)과 로버스트 설계 모형

반응표면분석(RSM : Response Surface Methodology)에 의해 반응치 $\boldsymbol{y}$와 CBFS에 의해 선택된 인자들 $\boldsymbol{x}$의 함수관계식을 추정하고자 한다. 그리고 추정된 함수식을 이용하여 공정의 평균과 분산을 추정한 다음, 로버스트 설계 모형을 세워 잡음인자에 둔감한 공정의 최적가동조건을 구하고자 한다.

반응표면분석은 여러 개의 인자 $x_1, x_2, x_3, \cdots, x_k$가 복합적인 작용을 함으로써 어떤 반응변수 y에 영향을 주고 있을 때, 이러한 반응의 변화가 이루는 반응표면에 대한 통계적인 분석방법을 말한다. 일반적으로 반응표면분석은 정확한 함수 관계를 알지 못하거나 또는 복잡할 때 입력치 $\boldsymbol{x}$와 반응치 $\boldsymbol{y}$의 함수 형태를 추정함으로써 이 반응치를 최적화하는데 사용되어진다.

본 연구에서는 Montgomery[13]가 제시한, 잡음인자와 제어인자를 모두를 통합한 반응 모형을 사용하고자 한다. k개의 제어인자 $\boldsymbol{x} = [x_1, x_2, \cdots, x_k]$와 r개의 잡음인자 $\boldsymbol{z} = [z_1, z_2, \cdots, z_r]$에 대한 추정된 반응함수는 식 (3)과 같다.

$$y(\boldsymbol{x}, \boldsymbol{z}) = f(\boldsymbol{x}) + h(\boldsymbol{x}, \boldsymbol{z}) + \psi \quad \cdots\cdots (3)$$

여기서 $f(\boldsymbol{x})$는 제어인자만 포함된 항으로 이루어진 부분이고, $h(\boldsymbol{x}, \boldsymbol{z})$는 제어 인자와 잡음인자 사이의 교호작용 항과 소음인자만의 항으로 이루어진 부분이다. ψ는 평균이 0이고 σ_z^2 분산을 가지는 정규분포를 가정한 오차 항이다.

추정된 반응함수식인 식 (3)의 기대치를 구함으로써 식 (4)와 같은 공정 평균에 대한 반응함수를 구할 수 있다.

$$E_z[y(\boldsymbol{x}, \boldsymbol{z})] = f(\boldsymbol{x}) \quad \cdots\cdots (4)$$

데일러 급수를 사용하여, 공정의 분산에 대한 반응함수를 구하면 식 (5)와 같다.

$$V_z[y(\boldsymbol{x},\boldsymbol{z})]=\sigma_z^2\sum_{i=1}^{r}\left[\frac{\partial h(\boldsymbol{x},\boldsymbol{z})}{\partial z_i}\right]+\sigma^2 \quad \cdots\cdots (5)$$

여기서 σ_z^2은 잡음인자의 분산을 나타내며, σ^2은 분산분석에서 평균제곱오차를 나타낸다.

로버스트 설계의 기본적인 개념은 변동에 의해 발생하는 영향력을 최소화함으로써 제품의 품질을 향상시키는 것이다. 이는 제품의 수행도가 다양한 원인의 변동에 덜 민감하도록 제품과 공정의 설계를 최적화함으로써 얻을 수 있다. 로버스트 설계는 제조비용과 개발시간의 감소와 높은 품질의 제품을 생산하기 위한 기술방법론을 말한다[16]. 로버스트 설계의 기본 개념은 최초로 다구찌에 의해 제시되었다. 다구찌 절차는 제품의 설계에서 제어인자의 최적 조건을 구함으로써 제품과 공정의 설계에서 품질을 개선시킬 수 있다고 강조하였다. 본 연구에서 사용하는 대표적인 로버스트 설계 모형 두 가지는 다음과 같다.

2.2.1 Dual Response Model

Vining and Myers(1990)가 제시한 dual response 모형은 실험 데이터를 기초로 공정 평균과 분산을 따로 분리해서 모델링함으로써, 식 (6)과 같이 공정 평균을 목표값으로 조절하고 분산을 최소화하여 로버스트 설계의 목적을 달성한 것이다.

$$\begin{aligned} &\text{Minimize} && \widehat{\sigma}(\boldsymbol{x}) \quad \cdots\cdots (6)\\ &\text{Subject to} && \widehat{\mu}(\boldsymbol{x})=\tau\\ & && \boldsymbol{x}\in\Omega \end{aligned}$$

여기서 τ와 Ω는 각각 품질 특성치의 목표값과 제어인자의 범위를 나타낸다.

2.2.2 MSE Model

Cho(1994)와 Lin and Tu(1995)은 zero-bias 논리를 기초로 한 최적화 계획은 추정된 평균을 강제로 지정된 값으로 두는 제약의 비현실성에 의해 오해하기 쉽다고 지적하였다. 그들이 제시한 최적화 모형을 식 (7)과 같다.

$$\begin{aligned} &\text{Minimize} && MSE=(\hat{\mu}(\boldsymbol{x})-\tau)^2+\hat{\sigma}^2(\boldsymbol{x}) \quad \cdots\cdots (7)\\ &\text{Subject to} && \boldsymbol{x}\in\Omega \end{aligned}$$

이 모형은 비록 공정의 평균이 목표값을 약간 벗어날지라도 MSE를 최소화하는 것은 보다 설득력 있는 해를 제공하다는 것을 가정한다.

3. 수치예제

본 연구에서 제시한 데이터 마이닝 기법을 이용한 로버스트 설계 모형의 최적화 절차를 구체적으로 설명하기 위하여 다음과 같은 수치예제를 제시하고자 한다. 사용된 데이터 집합은 담배 제조 공정에서 실시간으로 연속적으로 발생하는 데이터이다. 담배 제조공정은 원료가공공정, 궐련제조공정, 포장공정 3단계로 나누어진다. <표 1>은 원료 가공공정에서 얻어진 데이터이고, 반응치 3개와 제어인자 14개와 잡음인자 2개로 총 19개의 인자로 구성되어있다.

<표 1> 담배 원료가공공정의 데이터

No	y_1	y_2	y_3	x_1	⋯	z_1	z_2
1	1.55	20.05	1.38	2.02	⋯	16.84	66.56
2	1.63	12.58	2.64	2.62	⋯	16.01	64.57
3	1.66	18.56	1.56	2.08	⋯	15.72	63.41
4	1.52	18.56	2.22	2.20	⋯	17.17	62.34
5	1.70	14.02	2.85	2.38	⋯	15.35	64.24
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
200	1.71	22.10	1.52	2.23		14.74	67.07

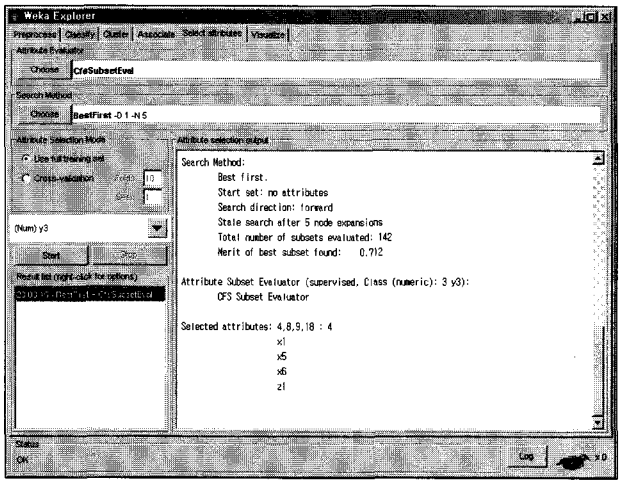
반응치에서 y_1은 연소속도, y_2는 당분함량, y_3는 니코틴 함량을 나타낸다. 제어인자는 x_1(질소), x_2(염소), x_3(칼륨), x_4(인), x_5(칼슘), x_6(마그네슘), x_7(휘발성 유기산), x_8(비휘발성 유기산), x_9(에테르 추출물), x_{10}(전환성 물질), x_{11}(pH), x_{12}(탄소), x_{13}(파네솔), x_{14}(프로필렌)로 구성되어 있다. 잡음인자로는 z_1(수분함량), z_2(건조온도)이 있다.

수치예제에서 반응 특성치를 니코틴 함량 y_3로 두고 나머지 18개 인자를 입력 인자로 두었다. 데이터 마이닝 기법 중 CBFS을 적용하기 위하여 데이터 마이닝 프로그램 중에 하나인 'Weka'을 사용하였다. Weka는 사용하기 편리하며 무료로 제공하고 있다. 많은 예제들을 제공하고 있으며 수많은 데이터 마이닝 알고리즘들이 내장되어 있다.

<그림 3>과 <표 2>는 Weka 프로그램 내에 있는 인자선택 알고리즘 CBFS와 BFS를 사용하여 반응치 y_3에 영향을 미치는 주요 인자 선택결과를 나타낸 것이다. 분석 결과, 니코틴 함량 y_3에 영향을 미치는 인자로 x_1(질소), x_5(칼슘), x_6(마그네슘)와 z_1(수분함량)이 선택되었다.

데이터 마이닝을 통해 얻은 주요인자들의 추정된 반응 함수를 얻기 위하여 반응표면분석을 실시하였다. 그 결과는 <그림 4>에서 보여주고 있다. F 검정과 p값을

통해 이차 회귀 모형이 유의함을 알 수 있고, R-sq 값이 87.3%로 이 모형이 반응함수로 사용하기 적합하다는 것을 의미한다.



<그림 3> CBFS와 BFS 분석결과

<표 2> CBFS와 BFS의 분석결과 정리

Selected Evaluator	The response attribute	y_3
	Merit of best subset	0.712
	Selected attributes	x_1, x_5, x_6, z_1
Search method	Search method	Best First
	Search Direction	forward
	Start set	no attributes
	Total number of subsets evaluated	142

Response Surface Regression: y versus x1, x2, x3, z1

The analysis was done using coded units.

Estimated Regression Coefficients for y

Term	Coef	SE Coef	T	P
Constant	2.1178	0.01587	133.435	0.000
x1	0.0541	0.02214	2.442	0.017
x2	0.0309	0.03703	0.833	0.407
x3	0.1324	0.02876	4.603	0.000
z1	0.1524	0.02105	7.241	0.000
x1*x1	0.2447	0.02986	8.196	0.000
x2*x2	-0.0347	0.04895	-0.708	0.481
x3*x3	0.1934	0.04424	4.372	0.000
x1*x2	-0.1320	0.06803	-1.941	0.056
x1*x3	-0.2080	0.04234	-4.912	0.000
x1*z1	-0.0710	0.03074	-2.309	0.023
x2*x3	0.0185	0.06710	0.276	0.784
x2*z1	0.0372	0.05649	0.658	0.512
x3*z1	-0.1529	0.02702	-5.660	0.000

S = 0.05501 R-Sq = 87.3% R-Sq(adj) = 85.4%

Analysis of Variance for y

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Regression	13	1.79566	1.795657	0.138127	45.65	0.000
Linear	4	1.13937	0.333010	0.083253	27.52	0.000
Square	3	0.38904	0.302771	0.100924	33.36	0.000
Interaction	6	0.26724	0.267239	0.044540	14.72	0.000

<그림 4> RSM 분석결과

RSM 분석결과의 회귀계수를 식 (3)에 대입하여 추정된 반응함수를 구하면 다음과 같다.

$$\begin{aligned} y(\boldsymbol{x},\boldsymbol{z}) = {} & 2.118 + 0.054x_1 + 0.031x_2 \\ & + 0.132x_3 + 0.152z_1 + 0.245x_1^2 - 0.035x_2^2 \\ & + 0.193x_3^2 - 0.132x_1x_2 - 0.208x_1x_3 0.019x_2x_3 \\ & - 0.071x_1z_1 + 0.037x_2z_1 - 0.153x_3z_1 \end{aligned}$$

식 (4)와 식 (5)을 이용하여 공정 평균과 분산의 추정된 반응 함수를 구하면 다음과 같다.

$$\begin{aligned} E_z[y(\boldsymbol{x},\boldsymbol{z})] = {} & 2.118 + 0.054x_1 + 0.031x_2 + 0.132x_3 \\ & + 0.245x_1^2 - 0.035x_2^2 + 0.193x_3^2 - 0.132x_1x_2 \\ & - 0.208x_1x_3 + 0.019x_2x_3 \end{aligned}$$

$$\begin{aligned} V_z[y(\boldsymbol{x},\boldsymbol{z})] = {} & 0.0204425[0.152 - 0.071x_1 + 0.037x_2 - 0.153x_3]^2 \\ & + 0.003026 \end{aligned}$$

추정된 공정 평균과 분산 식을 다음과 같이 dual response 모형과 MSE 모형에 적용시켰다.

[Dual Response Model]

Minimize $\quad \hat{\sigma}(\boldsymbol{x},\boldsymbol{z})$

Subject to $\quad \hat{\mu}(\boldsymbol{x},\boldsymbol{z}) = 0.8$

$\quad \boldsymbol{x} \in \Omega$

[MSE Model]

Minimiz $\quad MSE = (\hat{\mu}(\boldsymbol{x},\boldsymbol{z}) - 0.8)^2 + \hat{\sigma}^2(\boldsymbol{x},\boldsymbol{z})$ ····· (6)

Subject to $\quad \boldsymbol{x} \in \Omega$

<표 3>는 목표값 $\tau = 0.8$인 니코틴 함량 y에 대한 두 모형의 최적해를 구한 것이다. dual response 모형에서 평균은 목표값과 같으나 분산이 MSE 모형에 비해 크게 나왔다.

<표 3> $\tau = 0.8$일 때, 로버스트 모형의 최적해

Model	$\boldsymbol{x}$			$\hat{\mu}(\boldsymbol{x})$	$\hat{\sigma}^2(\boldsymbol{x})$
	x_1	x_2	x_3		
dual response	0.929	6.050	1.002	0.800	0.00353
MSE	1.704	6.068	1.632	0.795	0.00303

4. 결 론

자동화된 제조공정의 데이터 베이스로부터 얻은 데이터를 사용하여 공정을 실시간으로 관리하기 위해서는

방대한 양의 데이터들과 인자들을 관리할 수 있어야 한다. 본 연구에서는 이를 위하여 데이터 마이닝 기법을 사용하여 수많은 인자들 중에서 품질 특성치에 영향을 미치는 주요 인자를 선택하였다.

본 연구에서 제시한 기법은, 별도의 샘플링이나 실험계획법을 통해 데이터를 구하지 않고 현장의 데이터를 그대로 사용하여 공정을 관리할 수 있기 때문에 비용적인 면에서나 분석적인 면에서 많은 효율성을 가져다준다. 데이터 마이닝으로 선택된 인자들의 데이터들을 사용하여 반응표면분석을 실시한 결과, 선택된 인자들의 정확한 함수관계를 구할 수 있었다. 이를 통해 공정의 평균과 분산을 구하여 로버스트 설계 모형을 세워서 공정의 평균과 분산이 잡음인자에 로버스트 한 공정의 최적가동조건을 구하였다.

본 연구는 급속도로 변화하는 제조공정에서 공정과 품질을 관리하고 개선하는데 있어서 보다 실용적이면서 과학적인 새로운 기법으로 사료된다.

참고문헌

[1] 김만선, 이상용; "신경망을 이용한 대용량 데이터 처리를 위한 군집화 방법", 공주대학교 생산기술연구소 논문집, 10, 2002.

[2] 이현우, 남호수, 강중철; "A Study on Data Mining Application Problem in the TFT-LCD Industry", 한국데이터정보과학회지, 16(4) : 91-101, 2005.

[3] 이기훈; "데이터 마이닝에서 로버스트 통계적 기법의 도입", 산경논총, 18, 2000.

[4] Besharati, B., Luo, L., and Azarm, S.; "Multi-Objective Single Product Robust Optimization: An Integrated Design and Marketing Approach," *Journal of Mechanical Design*, 128(4) : 884-892, 2006.

[5] Chang X. F. and Xian F. W.; "Data mining techniques applied to predictive modeling of the knurling process," *IIE Transactions*, 36 : 253-263, 2004.

[6] Cho D. H., Ahn B. S., and Kim S. H.; "Prioritization of association rules in data mining : Multiple criteria decision approach," *Expert Syst. Appl.*, 29(4) : 867-878, 2005.

[7] Cho, B. R.; "Optimization Issues in Quality Engineering," Ph.D. Dissertation, School of Industrial Engineering, University of Oklahoma, 1994.

[8] DuMonuchel, W.; "Bayesian Data Mining in Large Frequency Tables With an Application to the Spontaneous Reportign System," *The American Statistician*, 53 : 177-202, 1999.

[9] Gardner, M. and Bieker, J.; "Data Mining Solves Though Semiconductor Manufacturing Problem. Conference on Knowledge Discovery," in Data Proceedings of the sixth ACMSIGKDD international conference on Knowledge discovery and data mining, New York, pp. 376-383, 2000.

[10] Hall, M. A.; "Correlation-based Feature Selection for Machine Learning," Waikato University, Department of Computer Science. Hamilton, New Zealand, 1998.

[11] John, G. H., Kohavi, R., and Pflager, P.; "Irrelevant Features and the Subset Selection Problem," *In Machine Learning:* Proceedings of the Eleventh International Conference, Morgan Kaufmann, 1994.

[12] Lin, D. K. J. and Tu, W.; "Dual response surface optimation," *Journal of Quality Technology*, 27 : 34-39, 1995.

[13] Montgomery D. C.; *Introduction to Statistical Quality Control*, 4th edn. John Wiley & Sons, New York, 2001.

[14] Pignatiello, J. J. and Ramberg, J. S.; "Discussion of off-line quality control, parameter design and Taguchi method," *Journal of Quality Technology*, 17(4) : 151-161, 1985.

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

[16] Shin, S. M. and Cho, B. R., "Bias-specified robust design optimization and its analytical solutions," *Computer & Industrial Engineering*, 48 : 129-140, 2005.

[17] Fayyad, U., piatetsky-Shapiro, G., and Smyth, P.; "The KDD Process for Extracting Useful Knowledge from Volumes of Data," *Communication of the ACM*, 39(11) : 27-34, 1996.

[18] Vining, G. G. and Myers, R. H.; "Combining Taguchi and response surface Philosophies: A dual response approach," *Journal of Quality Technology*, 22 : 38-45, 1990.

[19] Witten, I. W. H. and Frank, E.; *Data Mining: Practical Machines Learning Tools and Techniques*, 2nd edn Morgan Kaufmann, San Francisco, 2005.

[20] Yu, L. and Liu, H.; "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution," The Proceedings of the 20th International Conference on Machine Learning (ICML-03). Washington D. C., pp. 856-863, 2003.