데이터 마이닝과 통계적 기법을 통합한 최적화 기법

송서일* · 정혜진*† · 신상문**

* 동아대학교 산업경영공학과
** 인제대학교 시스템경영공학과

Optimization Methodology Integrated Data Mining and Statistical Method

Suh-Ill Song* · Hey-Jin Jung*† · Sang-Mun Shin**

* Dept. of Industrial & Management Systems Engineering, Dong-A University
** Department of Systems Management & Engineering, Inje University

Key Words : Data Mining, CBFS, SPC, RSM, Optimization

Abstract

These days manufacture technology and manufacture environment are changing rapidly. By development of computer and enlargement of technique, most of manufacture field are computerized. In order to win international competition, it is important for companies how fast get the useful information from vast data. Statistical process control(SPC) techniques have been used as a problem solution tool at manufacturing process until present. However, these statistical methods are not applied more extensively because it has much restrictions in realistic problems. These statistical techniques have lots of problems when much data and factors are analyzed.

In this paper, we proposed more practical and efficient a new statistical design technique which integrated data mining (DM) and statistical methods as alternative of problems. First step is selecting significant factor using DM feature selection algorithm from data of manufacturing process including many factors. Second step is finding optimum of process after estimating response function through response surface methodology(RSM) that is a statistical techniques

1. 서 론

컴퓨터의 발달과 기술력의 증대로 인해 우리 일상생활에서 뿐만 아니라 제조 공정 시스템이 자동화됨에 따라 하루에도 수 천 수억 개의 품질 특성치들이 계측된다. 이렇게 발생한 데이터들은 데이터베이스화 되어 실시간으로 공정의 상태를 파악하는데 사용된다. 이 데이터로부터 유용한 정보를 빨리 찾고 분석하여 제품 설계에 반영함으로써 공정의 문제점을 찾아내어 공정을 개선시킬 수 있다. 과거부터 지금까지 제조업에서 공정의 품질과 제품의 품질을 개선하기 위하여 통계적 공정관리(statistical process control : SPC)와 통계적 품질관리(statistical quality control : SQC) 기법들이 유용하게 사용되어지고 있다. 하지만 이러한 제조 환경에서 기존의 통계적 기법을 사용하여 공정을 관리하는 것은 현실적으로 많은 어려움이 발생한다.

통계적 품질관리 기법은 적은 양의 데이터를 정확하게 분석하지만, 수백만 또는 수 억 개의 데이터를 분석하는 것은 어렵다. 분석 자체도 힘들지만 표본의 크기가 커지면 '의미 없는' 차이도 유의하게 판정되는 오류도 발생 한다[2]. 그리고 대부분의 통계

† 교신저자 tears000@donga.ac.kr
※ 본 연구는 2006년도 동아대학교 학술연구비(공모과제)에 의하여 연구되었음.

적 기법은 사전에 모집단에 대한 분포를 가정하고 인자들 간의 독립성을 가정하여 분석한다. 하지만 실제 현장에서 분석하고자 하는 데이터들은 이러한 가정들을 만족하지 못하는 경우가 많아서 통계적 기법들이 적용하지 못하는 경우도 발생하고, 또는 무조건적인 가정에 의해 잘못된 결과를 도출하는 경우도 발생 한다.

본 연구에서는 이러한 통계적 기법들의 현실적인 문제를 해결하는 방안으로 데이터 마이닝 기법을 제안하고자 한다. 통계적 기법은 어떤 목적에 의해 수집된 자료들을 분석하는 제 1 차적 데이터 분석인데 반해 데이터 마이닝은 거대한 데이터베이스에서 관심이나 흥미를 가질 만한 숨겨진 관계를 찾아보는 제 2 차적인 데이터 분석 기법이다[1]. 하지만 데이터 마이닝 기법이 통계학 분야에서 각광받지 못하는 이유는 찾아낸 패턴들이 임의적인 현상일 수 있다는 불확실성과 수리적 모형과 해석의 어려움 때문이다. 그래서 본 연구에서는 이러한 통계적 기법들과 데이터 마이닝의 단점들은 보완하면서 장점들은 절충하는 새로운 통계적 설계 기법을 제시하고자 한다.

기존의 데이터 마이닝의 연구들은 크게 두 가지로 나눌 수 있는데, 하나는 적용 사례에 관한 연구이고 다른 하나는 데이터 마이닝의 알고리즘 개발 및 비교 평가 연구이다. 첫 번째 경우에는 다양한 분야에 데이터 마이닝 기법을 적용한 사례들에 관한 연구이다. 데이터 마이닝은 데이터의 특성에 따라 적용되어지는 기법들이 다양하다. 이미 입력과 결과가 결정되어 있는 관리된 데이터(supervised data)인 경우에는 입력과 결과 사이에 어떤 패턴관계가 있는가를 찾아내고 이를 바탕으로 미래의 결과를 예상함으로써 보다 효율적인 의사결정을 지원하기 위해 데이터 마이닝 기법이 적용된다[5]. 예를 들면, 금융기관의 개인 신용평가나 신용카드 회사의 사기감지, 통신 서비스회사의 고객 이탈 방지 등에 적용된다. 입력변수는 가지고 있지만 결과변수는 가지고 있지 않는 관리되지 않은 데이터(unsupervised data)인 경우에는 입력 변수들을 중심으로 데이터 사이의 연관성과 유사성을 찾아내기 위하여 데이터 마이닝 기법이 적용되어진다. 예를 들면, 소비자의 구매 패턴을 파악하여 매장의 제품을 진열하거나 제품이나 서비스의 교차 판매하여 상품의 매출 증대시키고, 제조업체에서의 불량 또는 결함을 관리하며, 병원에서의 질병 진단 등에 적용되어지는 사례들이 많다.

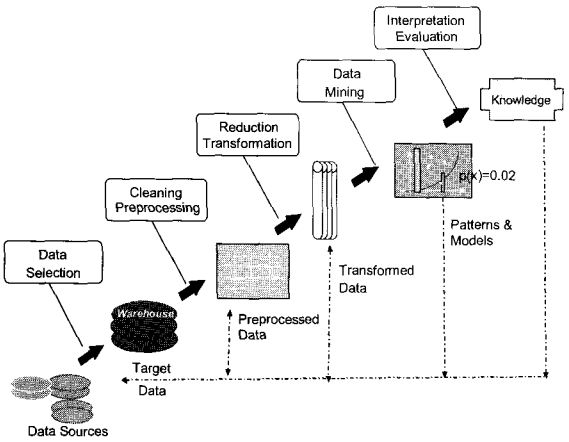
두 번째 경우에는 신경망 분석, 의사결정나무 분석, 동시발생 매트릭스, K-Means Clustering 등 많은 기존의 알고리즘을 비교 평가하거나 보다 향상된 새로운 알고리즘을 개발하여 제시하는 연구들이다[3]. 주로 CRM이나 사회 과학 분야에서 많이 연구되었던 데이터 마이닝 기법들이 최근에는 제조 공정에서도 많이 적용되어지고 있다[9]. Written and Frank(2000)는 반도체 공정에 데이터 마이닝 기법을 적용하였고[6] Feng and Wang(2002)은 knurling 공정의 예측 모형을 위에 데이터 마이닝 기법을 적용한 사례를 제시하였다[4]. 현재의 데이터 마이닝의 추세는 사회 과학과 제조 공정뿐만 아니라 컴퓨터 과학 분야인 동영상 및 멀티미디어 데이터의 마이닝 작업에 대한 연구가 많은 관심을 끌며 이루어지고 있는 추세이다. 하지만 기존의 데이터 마이닝 논문들은 여러 알고리즘을 사용하여 인자들 간의 관련성을 규명하는데만 초점을 두고 있다. 하지만 본 연구에서는 인자들 간의 관련성 규명하고 한 단계 더 나아가 이를 토대로 수리적인 모형을 구축하여 이 모형의 최적해를 제시하고자 한다.

본 연구는 크게 두 단계로 나누어 연구되어지고 있다. 첫 번째 단계에서는 제조공정 데이터에 데이터 마이닝 기법을 적용하여 많은 품질 특성치들 중에 반응치(output)에 영향을 미치는 유의한 인자(input)를 선택한다. 두 번째 단계에서는 통계적 분석 기법인 반응표면분석(RSM : response surface methodology)을 통하여 이 인자들의 유의성을 검토하고 반응 품질 특성치에 대한 추정식을 구한 다음, 공정의 최적 조건을 찾고자 한다.

2. 데이터 마이닝 기법과 알고리즘

데이터 마이닝(DM: Data Mining)은 대용량의 데이터로부터 유용하게 활용될 수 있는 지식을 효과적으로 찾아내는 지식 탐사의 한 연구 분야이다. 즉 유용한 정보의 추출을 위한 방법론이라고 할 수 있다. 종종 데이터마이닝과 지식발견(KDD, knowledge discovery in database)이라는 용어를 혼용해서 사용하고 있다. 데이터마이닝은 통계학자, 데이터분석가, 그리고 데이터베이스 분야에서 많이 사용되는 용어이고 지식발견(KDD)은 인공지능이나 기계학습(machine learning) 분야에서 자주 사용되는 용어이다. <그림 1>에서 보여주는 것처럼, 지식발견

은 데이터로부터 유용한 정보를 발견하는 전체 프로세스이고 데이터마이닝은 지식발견 프로세스 중에서 데이터로부터 정보를 추출하기 위하여 기법을 적용하는 특정단계로 정의하고 있다[14].



<그림 1> KDD 프로세스의 5단계

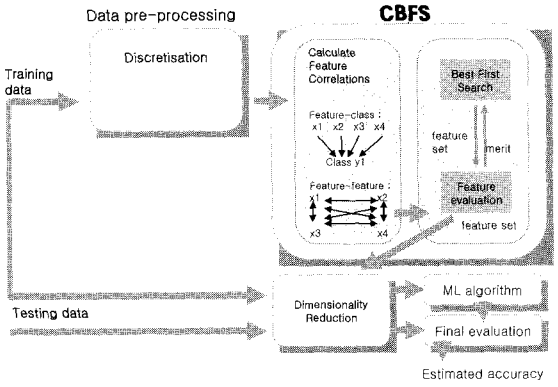
거대한 데이터베이스(DB) 공간으로부터 유의한 특성들을 선택하는 단계는 데이터 마이닝 과정에서 매우 중요한 단계이다. 특성선택(feature selection) 알고리즘은 filter와 wrapper 두 가지 기법이 있다[8]. filter 기법은 관련 없는 인자를 걸러주는 여과기와 같은 방법으로 다루어지고, wrapper 기법은 인자들의 평가 함수의 일부로써 귀납 알고리즘을 사용한다. 주요한 인자 집합을 평가하는 학습 알고리즘을 사용하는 wrapper 기법보다 데이터의 일반적인 특성을 기초로 한 휴리스틱 방법을 사용하는 filter 기법이 보다 빠르며 높은 차원의 데이터에 대하여 보다 실용적이기 때문에 더 선호한다[15]. 본 연구에서는 방대한 제조공정의 데이터로부터 반응치에 영향을 미치는 인자들을 선택하여 통계적 분석을 하고자 함으로 filter 기법에 하나인 CBFS 기법을 사용하고자 한다.

CBFS(Correlation-based Feature Selection)는 휴리스틱 평가함수를 기초로 한 상관관계에 따라 입력 특성(feature)의 부분 집합에 대하여 순위를 매기는 filter 알고리즘이다[7]. 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 다음 식 (1)과 같다.

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

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.

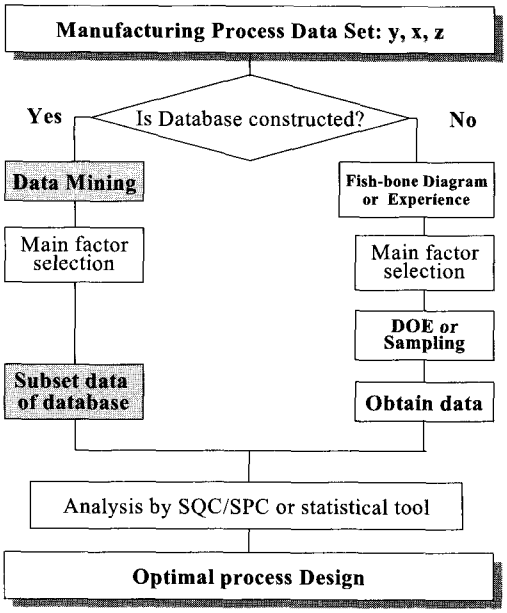
모두를 열거하는 방법을 사용하여 가장 좋은 부분집합을 구하는 것은 거의 불가능하다. 수많은 부분집합을 평가하기 위해 탐색공간을 줄이는 가장 효율적인 방법 중에 하나가 BFS(Best First Search) 기법이다. 이 탐색 방법은 위에 있는 CBFS 알고리즘을 수행하기 위한 휴리스틱 탐색기법이다. 이는 탐색 경로를 따라 뒤로 돌아가는(backtracking)것을 허용하는 개선된 탐색 전략이다. 만약 탐색하던 경로가 가망성이 부족해 보이면, BFS는 가망성이 보다 높아 보이는 이전의 부분집합으로 되돌아간다.



<그림 2> CBFS와 BFS 알고리즘

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. <그림 2>는 위에서 설명한 CBFS와 BFS 알고리즘 절차를 나타낸 것이다.



<그림 3> 제조공정에서 데이터 분석절차

본 연구는 <그림 3>과 같이 자동화된 제조공정에서 데이터베이스가 구축되어있는 경우에 데이터 마이닝 기법을 통해 반응치 y에 영향을 미치는 인자를 선택하고자 한다.

과거에는 주요인자들 선택하기 위하여 전문가의 의견이나 또는 과거의 경험에 의해 선택하였다. 하지만 이렇게 주관적으로 인자를 선택하는 것보다는 실제 현장의 데이터를 사용하여 인자를 선택하는 것이 보다 객관적이고 인식하지 못했던 인자들까지도 고려할 수 있다. 뿐만 아니라 별도의 시간과 비용이 드는 샘플링이나 실험을 통해 데이터를 얻을 필요없이 데이터베이스로부터 선택된 인자들의 데이터를 그대로 사용하여 기존의 통계적 기법들을 통해 최적공정설계를 하고자 한다. 따라서 데이터 마이닝 특성선택 알고리즘을 사용함으로써 객관성뿐만 아니라 실용성 면에서도 효율성을 얻을 수 있다.

3. 통계적 기법과 최적화

데이터 마이닝 특성선택 알고리즘을 통해 거대한 데이터베이스 공간으로부터 반응치와 상관이 높은 인자들을 선택하였다. 그리고 이 인자들의 수리적 모형을 세워 공정의 최적 조건을 구하기 위하여 본 연구에서는 통계적 기법 중에 하나인 반응표면분석(RSM : response surface methodology)을 적용하였다.

반응표면분석(RSM)은 반응치가 몇몇 입력 요인에 의해 영향을 받는 경우에 이를 모형화하고 분석하는데 매우 유용한 도구이다. 일반적으로 RSM은 정확한 함수 관계를 알지 못하거나 또는 복잡할 때 입력치와 반응치의 함수 형태를 추정함으로써 이 반응치를 최적화하는데 사용되어진다. RSM은 실험계획법, 모형 적합도와 최적화의 부분에서 적용되어진다[10, 11].

반응치(response) y와 관련된 입력 인자 x을 사용하여 반응치 함수($\hat{y}(x)$)를 추정하면 다음에 오는 식(2)과 같다.

$$\hat{y}(x)=\hat{\beta}_0+\sum_{i=1}^{k}\hat{\beta}_i x_i+\sum_{i=1}^{k}\hat{\beta}_{ii}x_i^2 +\sum_{i}\sum_{i<j}^{k}\hat{\beta}_{ij}x_i x_j \quad (2)$$

여기서 β들은 인자 x_i에 대해 추정된 회귀 계수들이다. $\hat{\beta}_0$는 상수항의 계수이고, $\hat{\beta}_i$는 일차항의 계수이고, $\hat{\beta}_{ii}$는 이차항의 계수이며, $\hat{\beta}_{ij}$는 x_i와 x_j의 교호작용의 계수이다. RSM에서 중요한 점은 추정된 반응치 함수가 통계적으로 유의한가를 확인해야 한다. 분산분석(ANOVA)을 통해 추정된 반응치 함수의 유의성을 평가하고 이렇게 추정된 반응치 함수는 공정 모수의 최적화를 위해 사용되어진다. 반응표면분석을 통해 얻어진 추정된 식을 이용하여 공정의 최적화 모형을 세우면 식 (3)과 같다.

$$\begin{aligned} &\text{Minimize} && [\hat{y}(x)-\tau]^2 \\ &\text{Subject to} && x\in\Omega \end{aligned} \quad (3)$$

여기서 τ는 목표값을 나타낸다. 추정된 식과 목표값 τ의 차이가 최소화되는 주요인자들의 최적해를 구하고자 한다.

4. 수치예제

입력 인자 사이에는 반응치 y에 매우 영향을 미

치는 인자들도 있고 그렇지 않은 인자들도 있다. 실제 반도체 공정이나 화학 공정에서는 수많은 인자들을 포함하는 경우가 많이 있다. 이렇게 많은 인자들로부터 반응치에 영향을 미치는 인자를 선택하는 것은 쉬운 일이 아니다. 본 연구에서는 데이터 마이닝 기법을 통해 반응치에 영향을 미치는 주요 인자를 선택하고자 한다.

본 연구에서 사용되어지는 데이터의 집합은 담배 제조 공정에서 실시간으로 연속적으로 발생하는 데이터이다. 본 연구에서 제시한 기법을 설명하고자 하는 목적이므로 실제 현장의 데이터를 약간 수정하고 요약하였다. 담배 제조공정은 원료가공공정, 궐련제조공정, 포장공정 3단계로 나누어진다. 수치예제에서 사용한 데이터는 원료가공공정으로부터 얻어진 데이터이다. 원료가공 공정에서는 여러 종류의 잎담배를 제품별 설정된 표준에 따라 투입하여 가습, 가향, 조화 등 제품특성에 맞게 가공되어진다.

<표 1> 담배 원료가공공정의 데이터

No	y1	y2	y3	x1	x2	x3	…
1	1.55	20.05	1.38	2.02	2.90	2.17	…
2	1.63	12.58	2.64	2.62	2.78	1.72	…
3	1.66	18.56	1.56	2.08	2.68	2.40	…
4	1.52	18.56	2.22	2.20	3.17	2.06	…
5	1.70	14.02	2.85	2.38	2.52	2.18	…
6	1.68	15.64	1.24	2.03	2.56	2.57	…
7	1.78	14.52	2.86	2.87	2.67	2.64	…
8	1.57	18.52	2.18	1.88	2.58	2.22	…
9	1.60	17.84	1.65	1.93	2.26	2.15	…
10	1.52	13.38	3.28	2.57	1.74	1.64	…
11	1.68	17.55	1.56	1.95	2.15	2.48	…
12	1.74	17.97	2.00	2.03	2.00	2.38	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
800	1.71	22.10	1.52	2.23	3.20	2.39	

<표 1>은 원료 가공공정에서 얻어진 데이터이고 반응치 3개와 인자 16개로 총 19개의 인자로 구성되어있다. 반응치에서 y_1은 연소속도, y_2는 당분함량, y_3는 니코틴 함량을 나타낸다. 인자는 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), x_{15}(규소), x_{16}(납)으로 구성되어 있다. 수치예제에서는 반응 특성치를 니코틴 함량 y_3로 두고 니코틴 함량 y_3에 영향을 미치는 인자들을 선택하고자 한다.

<표 2> CBFS 분석결과

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

데이터 마이닝 기법을 적용하기 위하여 Weka 프로그램 사용하였다. CBFS와 BFS 알고리즘을 사용하여 주요 인자를 선택한 결과는 다음에 오는 <표 2>와 같이 나타났다. BFS 탐색기법을 사용하여 총 142개의 특성 집합을 평가하여 평가함수 값이 0.712로 가장 크게 나온 특성집합(x_1, x_5, x_6)를 선택하였다.

Response Surface Regression: y versus x1, x2, x3
The analysis was done using coded units.
Estimated Regression Coefficients for y

Tern	Coef	SE Coef	T	P
Constant	3.853	3.72936	1.033	0.014
x1	-9.770	2.22099	-4.399	0.000
x2	1.882	0.36774	5.117	0.000
x3	-2.092	2.50366	-0.836	0.026
x1*x1	4.112	0.55767	7.373	0.000
x2*x2	0.052	0.02879	1.803	0.055
x3*x3	1.636	0.49042	3.335	0.001
x1*x2	-0.942	0.21590	-4.365	0.000
z1*x3	1.239	0.55631	2.227	0.028
x2*x3	-0.531	0.12157	-4.368	0.000

S=0.06869 R-sq=81.3% s-Sq(adj) = 77.3%

Analysis of Variance for y

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Regression	9	1.63117	1.631174	0.181242	38.41	0.000
Linear	3	1.08705	0.208257	0.069419	14.71	0.000
Square	3	0.40644	0.345118	0.115039	24.38	0.000
Interaction	3	0.13769	0.137668	0.045896	9.73	0.000
Residual Error	90	0.42469	0.424690	0.004719		
Lack-of-Fit	89	0.42469	0.424690	0.004772	*	*
Pure Error	1	0.00000	0.000000	0.000000		
Total	99	2.05586				

<그림 4> RSM 분석결과

다음은 이 인자들의 통계적 분석과 공정의 최적화를 위하여 반응표면분석(RSM)을 시행하였다. 반응표면분석(RSM)을 통하여 위에서 구한 세 개의 입력 인자와 반응치의 회귀 추정식을 구하고 분산분석을 통하여 이 추정식의 유의한가를 알아보고자 한다. 인자 x_1, x_5, x_6은 편의상 x_1, x_2, x_3로 표시하였다. <그림 4>는 반응표면분석(RSM) 결과이다.

반응치에 대한 추정된 회귀 계수 β들을 식 (2)에 대입하여 다음과 같은 회귀 추정식을 구하였다.

$$\begin{aligned}\hat{y} = {} & 3.853 - 9.770x_1 + 1.882x_2 - 2.092x_3 \\ & + 4.112x_1^2 + 0.052x_2^2 + 1.636x_3^2 \\ & - 0.942x_1x_2 - 1.239x_1x_3 - 0.531x_2x_3\end{aligned}$$

F검정과 p값을 통해 이차 회귀모형이 유의함을 알 수 있고 R-sq 값이 81.2%로 이 모형의 반응함수로 사용하기 적합하다는 것을 의미한다. 니코틴 함량 y의 목표값은 2.1이므로 위의 추정식을 사용하여 최적 모형을 구하면 다음과 같다.

$$\begin{aligned}&\text{Minimize} && [\hat{y}(\boldsymbol{x}) - 2.1]^2 \\ &\text{Subject to} && \boldsymbol{x} \in \Omega\end{aligned}$$

이때 추정식 $\hat{y}(\boldsymbol{x})$와 목표값 τ의 차이의 제곱을 최소화하는 공정의 최적조건을 다음과 같이 구하였다.

$$x_1 = 1.1079,\ x_2 = 6.0360,\ x_3 = 0.5010$$

니코틴 함량이 목표값에 가장 가깝도록 공정의 최적가동조건을 구할 수 있다. 이뿐만 아니라 동일한 방법으로 나머지 반응치인 연소속도(y_1), 당분함량(y_2)에 대하여 데이터 마이닝 기법을 사용하여 주요 인자를 선택하고 반응표면분석을 통해 추정식을 구한 다음 공정의 최적가동조건을 구할 수 있다.

5. 결 론

제조 환경은 급속도로 변화되고 있다. 대부분의 시스템이 자동화되어 품질 특성치들도 자동으로 계측되고 매 시간마다 업데이트 된 데이터베이스로부터 데이터를 얻을 수 있다. 데이터 양이 방대하고 많은 인자들을 포함하고 있기 때문에 기존의 통계적 기법으로는 해결하기 어려운 문제들이 발생한다. 따라서 본 연구에서는 이러한 문제를 해결하는 하나의 방안으로 데이터 마이닝 기법과 통계적 기법를 통합하여 보다 실용적이고 효율적인 최적공정설계 기법을 제시하였다.

기존에 연구되어진 데이터 마이닝 기법처럼 인자간의 관련성을 규명하는데 그치는 것이 아니라 데이터 마이닝 알고리즘(CBFS)을 통해 상관이 높은 인자를 선택하고, 이 인자들이 통계적으로 유의한가를 정확하게 알아본 다음, 수리적인 모형을 세워서 최적조건을 구함으로써 보다 발전된 기법을 개발하였다.

본 논문에서 보여주고 있는 수치예제는 가장 단순한 사례에 불과하지만, 이러한 공정보다 더 많은 인자와 데이터양을 처리하거나 분석하는 화학공정이나 반도체 공정에서는 전통적인 통계적 기법보다는 본 연구에서 제시하는 데이터 마이닝을 이용한 새로운 통계적 기법이 유용한 도구로써 사용되어질 수 있을 것이다.

참 고 문 헌

[1] 백동현, 한창희(2003), "데이터마이닝을 이용한 반도체 FAB 공정의 수율개선 및 예측", 「한국지능정보시스템학회논문지」, 9권, 1호, pp. 157-177,

[2] 이기훈(2000), "데이터 마이닝에서 로버스트 통계적 기법의 도입", 「산경논총」, 18집.

[3] Alexander, H. and Daniel, A. K.(1999), "Clustering Method for Large Data Sets", SIGMOD99.

[4] Chang, X. F. and Xian, F. W.(2004), "Data mining techniques applied to predictive modeling of the knurling process", *IIE Transactions*, Vol. 36, pp. 253-263.

[5] DuMonuchel, W.(1999), "Bayesian Data Mining in Large Frequency Tables With an Application to the Spontaneous Reportign System", *The American Statistician*, Vol. 53, pp. 177-202.

[6] Gardner, M. and Bieker, J.(2000), "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.

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

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

[9] Kuralmani, V. and Xie, M.(2002), "A conditional decision procedure for high yield process", *IIE Transaction,* Vol. 34, pp. 1021-1030.

[10] Lin, D. K. J. and Tu, W.(1995), "Dual response surface optimation", *Journal of Quality Technology*, Vol. 27, pp. 34-39.

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

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

[13] Fayyad, U. and piatetsky-Shapiro, G. and Smyth, P.(1996), "The KDD Process for Extracting Useful Knowledge from Volumes of Data", *Communication of the ACM*, Vol. 39, No. 11, pp. 27-34.

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

[15] Yu, L. and Liu, H.(2003), "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.