데이터 마이닝과 통계적 기법을 통합한 최적화 기법

정혜진, 송서일
동아대학교 산업경영공학과

- Optimization Methodology Integrated Data Mining and Statistical Method -

Suh-Ill Song, Hey-Jin Jung
Dept. of Industrial & Management Systems Engineering, Dong-A University

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

Abstract

Nowaday manufacture technology and manufacture environment are changing rapidly. By development of computer and enlargement of technique, most of manufacture field are computerized. It is measured automatically do much quality characteristics thereby and great many data happen in a day. corporations is important if have gotten fast information that are useful from wide data to go first in international competition according to these change. Statistical process control(SPC) techniques are used as a problem solution tool at manufacturing process until present. However, this statistical methods is not applied more extensively because have much restrictions in realistic problem.

In this paper, wish to develop more realistic and scientific new statistical design techniques doing to integrate data mining(DM) and statistical methods by the alternative to cope these problem. First step selects significant factor using DM techniques from datas of manufacturing process including much factors and second step wish to find optimum of process after get the estimated response function through response surface methodology(RSM) that is statistical techniques

1. 서 론

컴퓨터의 발달과 기술력의 증대로 인해 우리 일상생활에서 뿐만 아니라 제조 현장에도 실시간으로 많은 양의 데이터들이 쏟아진다. 이러한 추세에 맞춰 급속도로 변하는 국가 또는 기업의 경쟁력에서 살아남는 위해서는 실제 현장 데이터로부터 유용한 정보를 빨리 찾고 분석하여 제품 설계에 반영하는 것이다. 과거부터 지금까지 제조업에서 공정의 품질과 제품의 품질을 개선하기 위하여 통계적 공정관리와 통계적 품질관리 기법이 유용하게 사용되어지고 있다. 하지만
방대한 데이터와 많은 인자를 포함하는 경우에
는 기존의 통계적 기법으로는 해결할 수 없는 많은 문제점들이 발생한다.

자동화 된 제조 공정에서는 많은 품질 특성치들이 자동으로 계측되고, 이들 데이터들이 데이터베이스(DB)화 되어 실시간으로 공정의 상태를 파악한다. 기존의 통계적 기법을 사용하여 수많은 품질 특성치로부터 공정을 관리하는 것은 많은 어려움이 발생한다. 통계적 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 발생한다. 그리고 대부분의 통계적 기법은 iid 정규성을 가정한지만 실제 현장에서 분석하고자 하는 데이터들은 이러한 가정을 만족하지 못해서 통계적 기법들이 적용되는 못하는 경우도 발생한다.

본 연구에서는 이러한 통계적 기법들의 현실적인 문제를 해결하는 방안으로 데이터 마이닝 (DM: Data Mining)기법을 제안하고자 한다. 통계적 기법은 어떤 목적에 의해 수집된 자료들을 분석하는 제 1 차적 데이터 분석인데 반해 데이터 마이닝은 거대한 데이터 베이스에서 관심이나 흥미를 가질 만한 숨겨진 관계를 찾아보는 제 2 차적인 데이터 분석 기법이다. 하지만 데이터 마이닝 기법이 통계학 분야에서 각광받지 못하는 이유는 찾아낸 패턴들이 임의적인 현상일 수 있다는 불확실성 때문이다. 그

래서 본 연구에서는 이러한 통계적 기법들과 데이터 마이닝의 단점들은 보완하면서 장점들은 절충하는 새로운 통계적 설계 기법을 제시하고자 한다.

기존의 데이터 마이닝(DM)의 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 마이닝을 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[4]. 예를 들면, 금융기관의 개인 신용평가나 신용카드 회사의 사기감지, 통신 서비스회사의 고객 이탈 방지 등에 적용된다. 입력변수는 가지고 있지만 결과변수(target)는 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 교차판매하여 상품의 매출 증대시키고, 제조업체에서의 불량 또는 결함을 관리하며, 병원에서의 질병 진단 등에 적용되어지는 사례들이 많다.

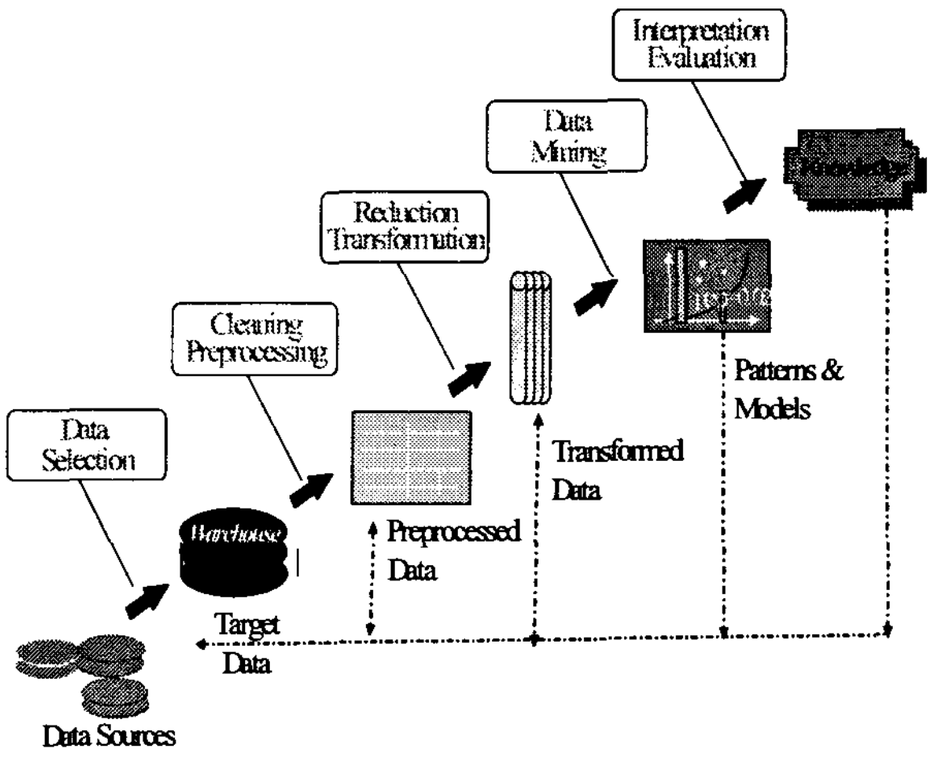
두 번째 경우에는 신경망 분석, 의사결정나무 분석, 동시발생 매트릭스, K-Means Clustering 등 많은 기존의 알고리즘을 비교 평가하거나 보다 향상된 새로운 알고리즘을 개발하여 제시하는 연구들이다[1]. 주로 CRM이나 사회 과학 분야에서 많이 연구되었던 데이터 마이닝 기법들이 최근에는 제조공정에서도 많이 적용되어지고 있다. Written and Frank(2000)는 반도체 공정에 DM 기법을 적용하였고[5] Feng and Wang(2002)은 knurling 공정의 예측 모형을 위에 DM 기법을 적용한 사례를 제시하였다[3]. 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 DM 논문들은 여러 알고리즘을 사용하여 인자들 간의 관련성을 규명하는데만 초점을 두고 있다. 하지만 본 연구에서는 인자들 간의 관련성 규명하고 한 단계 더 나아가 이를 토대로 모형을 구축하여 이 인자들의 최적해를 구하고자 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 선택한 다. 두 번째 단계에서는 통계적 분석 기법인 반응표면분석(RSM: response surface methodology)를 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. 따라서 데이터 마이닝을 효율적으로 수행하기 위해서는 시계열 분석 등 각종 통계적 기법과 데이터베이스 기술뿐만 아니라 신경망, 인공지능, 전문가시스템, 퍼지논리, 패턴인식 등과 같은 각종 정보기술과 기법들을 사용하게 된다. 종종 데이터 마이닝과 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. 지식발견은 데이터로부터 유용한 정보를 발견하는 전체 프로세스이고 데이터마이닝은 지식발견 프로세스 중에서 데이터로부터 정보를 추출하기 위하여 기법을 적용하는 특정단계로 정의하고 있다[14].

<그림 1>과 같이, KDD 프로세스는 5단계로 구성된다. 첫 번째 단계에서는, 지식발견의 목표와 문제 대상을 명확하게 정의한 후에, 데이터 집합 또는 변수 집합을 정의한다(Data Selection). 두 번째 단계는 사전처리(Preprocessing) 단계로써, 데이터의 적재, 변환, 클린징(cleaning)을 하는 단계이다. 특히, 데이터 클린징은 데이터 내의 잡음 제거, 이상치 데이터 필드에 처리 등에 관한 기본적인 작업을 수행하는 것으로 지식발견 프로세스 중에서 많은 시간과 노력을 요구하는 단계이다. 세 번째 단계는 데이터 변형(Transformation) 단계이다. 이 단계는 고려 대상이 되는 변수의 수를 줄이거나(reduction) 좀 더 효과적인 표현을 발견하여 변형시키는 단계이다. 네 번째 단계는 지식발견의 목표를 가장 효과적으로 달성할 수 있는 데이터마이닝 방법 또는 알고리듬이 무엇인지 선택하여 데이터마이닝을 수행하는 단계이다. 분류, 군집, 회귀분석 등 다양한 방법을 적용하여 데이터 속에 숨어 있는 패턴을 발견한다. 마지막 다섯 번째 단계는, 발견된 패턴을 사용자가 해석(interpretation) 가능하게 시각화하고 정의된 평가 기준에 의해 데이터 마이닝 결과를 평가(evaluation)한다[16][17]. 본 연구에서는 데이터 마이닝을 수행하는 단계에 대해서만 다루고자 한다.

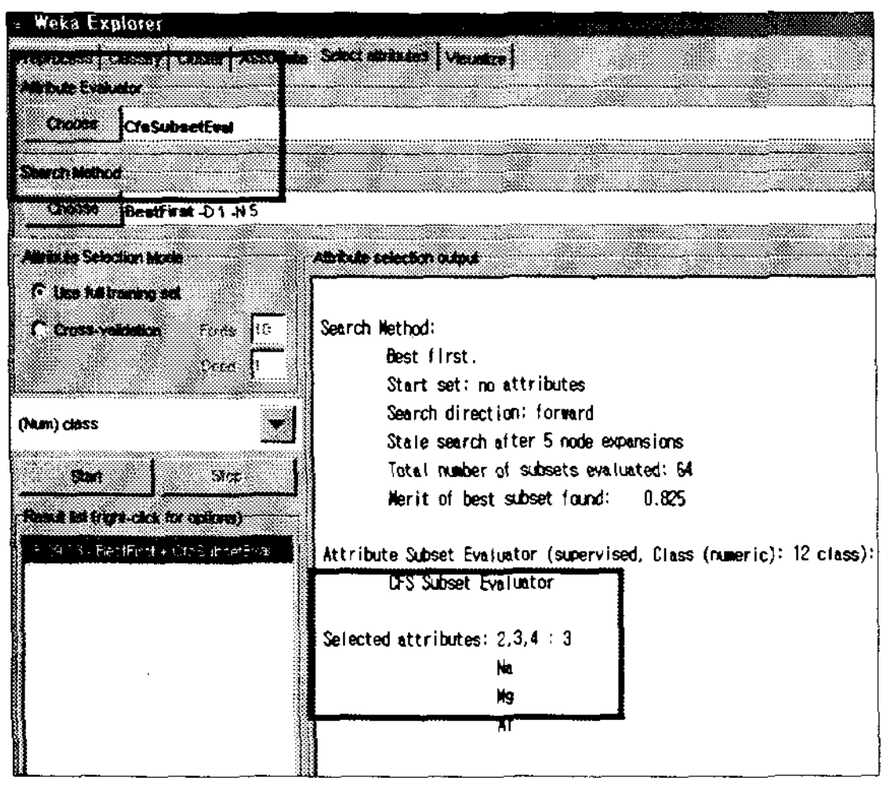


<그림 1> KDD 프로세스의 5단계

거대한 데이터베이스(DB) 공간으로부터 유의한 인자들을 선택하는 단계는 데이터 마이닝 과정에서

매우 중요한 단계이다. 인자 선택 알고리즘은 filter approach와 wrapper approach 두 가지가 있다[7]. filter approach는 관련 없는 인자를 걸러주는 여과기와 같은 방법으로 다루어지고, wrapper approach은 인자들의 평가 함수의 일부로써 귀납 알고리즘을 사용한다. 주요한 인자 집합을 평가하는 학습 알고리즘을 사용하는 wrapper approach보다 데이터의 일반적인 특성을 기초로 한 휴리스틱 방법을 사용하는 filter approach가 보다 빠르며 높은 차원의 데이터에 대하여 보다 실용적이기 때문에 더 선호한다. 본 연구에서 사용되는 데이터 마이닝 프로그램 "Weka"에서도 filter approach을 이용한 CBFS(Correlation-Based Factor Selection) 기법을 사용하고자 한다. <그림 2>에서 보여주는 것과 같이 "Weka" 프로그램은 자바 형식으로 작성되어있으며 소스 코드까지 공개를 하고 있어서 유용하게 널리 사용되어지고 있는 데이터 마이닝 프로그램이다[18].

Weka 프로그램에서는 두 가지 알고리즘을 사용하고 있다. 첫 번째 알고리즘은 CBFS (Correlation-Based Factor Selection)이다. CBFS는 휴리스틱 평가 함수를 기초로 한 상관관계에 의해 입력 요인들의 부분집합들 분류하는 filter approach 형식의 알고리즘이다[6][19]. 이 알고리즘의 평가 함수는 품질 특성치의 반응치에 매우 상관이 있는 인자뿐만 아니라 서로 상관이 없는 모든 인자를 다 포함한 부분집합에 대하여 행해진다. 인자들 사이에서도 주어진 반응치와 상관이 없거나 상관관계가 낮은 인자들은 무시되고, 비록 높은 상관관계가 있다하더라도 중복된 요인들은 제거한다. 두 번째 알고리즘은 BFS(Best First Search)이다. BFS는 CBFS 알고리즘을 수행하기 위한 휴리스틱 탐색 기법이다. 가장 좋은 인자 집합을 찾기 위하여 전부를 열거하는 방법은 많은 산업 현장에서 거의 불가능하다. BFS는 많은 인자 집합을 평가하는데 있어서 탐색 공간을 줄이는 가장 좋은 방법 중에 하나이다. 이 방법은 탐색 공간 경로를 따라 되돌아갈 수 있는 향상된 탐색 기법이다. 만약 탐색한 경로가 덜 유망해 보이면, 보다 더 유망해 보이는 이전 부분 집합으로 되돌아가서 탐색을 계속한다.



<그림 2> CBFS와 BFS 알고리즘을 이용한 Weka 프로그램

3. 통계적 기법과 최적화

데이터 마이닝을 통해 거대한 데이터 베이스 공간으로부터 반응치와 상관이 높은 인자들을 선택하였다. 이 인자들의 통계적 분석과 공정의 최적 조건을 구하기 위하여 본 연구에서는 통계적 기법 중에 하나인 반응표면분석(RSM: response surface methodology)을 다루고자 한다.

반응표면분석(RSM)은 반응치가 몇몇 입력 요인에 의해 영향을 받는 경우에 모형화 하고 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 내용에서 검토되어진다[13].

반응치(response) y와 관련된 입력 인자 x을 사용하여 반응치 함수($\hat{y}(x)$)를 추정하면 다음에 오는 식(1)과 같다.

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

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 확인해야한다. 분산분석(ANOVA)을 통해 추정된 반응치 함수의 유의성을 평가하고 이렇게 추정된 반응치 함수는 공정 모수의 최적화를 위해 사용되어진다. 공정의 최적화 모형은 식(1)와 같다.

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

여기서 τ는 목표값을 나타낸다.

4. 수치예제

입력 인자 사이에는 반응치 y에 매우 영향을 미치는 인자들도 있고 그렇지 않은 인자들도 있다. 본 연구의 수치예제에서는 입력 인자들이 그렇게 많지는 않지만, 실제 반도체 공정이나 화학 공정에서는 수많은 인자들을 포함하는 경우가 많이 있다. 이렇게 많은 인자들로부터 반응치에 영향을 미치는 인자를 선택하는 것은 쉬운 일이 아니다. 이런 경우에는 일반적인 통계적 기법으로도 해결할 수 없는 경우가 많다. 만약 인자들 선택하기 위하여 전문가의 의견이나 또는 과거의 경험에 의해 이루어진다면 공정에 영향을 미치는 중요 인자들이 빠질 수도 있으며 반대로 중요하지 않은 인자들이 선택되는 경우도 발생하게 된다. 그래서 본 연구에서는 반응치 y에 영향을 미치는 인자를 선택하기 위하여 데이터 마이닝 기법을 적용하였다.

수치예제에서 사용되고 있는 데이터는 유리 제조공정 과정에서 연속적으로 얻어진 데이터들이다. <표 1>에서 보여주는 것과 같이, 반응치 y에 영향을 미치는 인자들은 12개이고 400개의 데이터를 사용하여 분석하였다. 반응치 y와 입력 인자들에 대한 설명은 다음과 같다.

<표 1> 유리제조공정의 데이터 집합

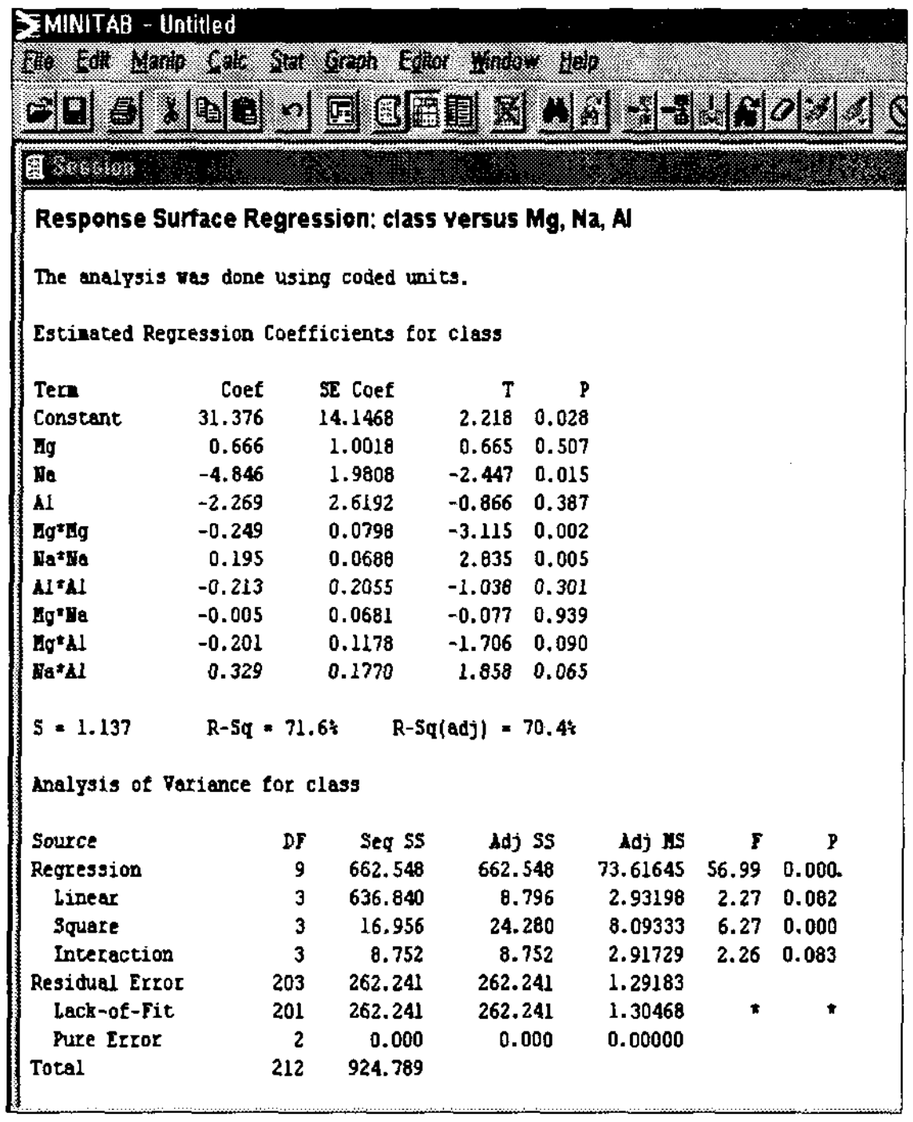
No	RI	Na	Mg	Al	Si	K	⋯	class
1	1.52	13.21	0.00	1.54	72.99	0.00	⋯	1
2	1.52	13.64	4.49	1.10	71.78	0.06	⋯	1
3	1.52	13.89	3.6	1.36	72.73	0.48	⋯	1
4	1.52	13.53	3.55	1.54	72.99	0.39	⋯	1
5	1.52	13.21	3.69	1.29	72.61	0.57	⋯	1
6	1.52	13.27	3.62	1.24	73.08	0.55	⋯	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋯	⋮
397	1.52	14.37	0.00	2.74	72.85	0.00	⋯	7
398	1.52	14.14	0.00	2.88	72.61	0.08	⋯	7
399	1.52	14.92	0.00	1.99	73.06	0.00	⋯	7
400	1.52	14.36	0.00	2.02	73.42	0.00	⋯	7

[입력인자 : x]
1. RI : 굴절지수(refractive index)
2. Na : 나트륨(Sodium)
3. Mg : 마그네슘(Magnesium)
4. Al : 알루미늄(Aluminum)
5. Si : 규소(Silicon)
6. K : 칼륨(Potassium)
7. Ca : 칼슘(Calcium)
8. Ba : 바륨(Barium)
9. Fe : 철(Iron)
10. Pb : 납(Lead)
11. B : 붕소(Boron)

[반응치 : y]
class(반응치 y) : 유리의 유형(Type)
1 - building windows float processed
2 - building windows non float processed
3 - vehicle windows float processed
4 - vehicle windows non float processed
5 - containers
6 - tableware
7 - head lamps

<그림 2>에서 보여주는 것과 같이 데이터 마이닝 기법을 적용하기 위하여 Weka 프로그램 사용하였다. 많은 인자들 중에 영향을 미치는 주요 인자를 선택하기 위하여 CBFS (Correlation-Based Factor Selection) 알고리즘을 사용하였고 탐색 기법으로는 BFS(Best First Search) 알고리즘을 사용하였다. Weka 프로그램 결과, 11개의 인자들 중에 반응치 y에 유의하게 영향을 미치는 인자로 Na, Mg, Al이 선택되었다.

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.



<그림 3> RSM 분석결과

분석 결과는 <그림 3>에서 보여주는 것과 같다. 반응치에 대한 추정된 회귀 계수 β 식(1)에 대입하여 다음과 같은 회귀 추정식을 구하였다.

$$\hat{y}(x) = 31.376 + 0.666x_1 - 4.846x_2 - 2.269x_3 - 0.249x_1^2 + 0.195x_2^2 - 0.213x_3^2 - 0.005x_1x_2 - 0.201x_1x_3 + 0.329x_2x_3$$

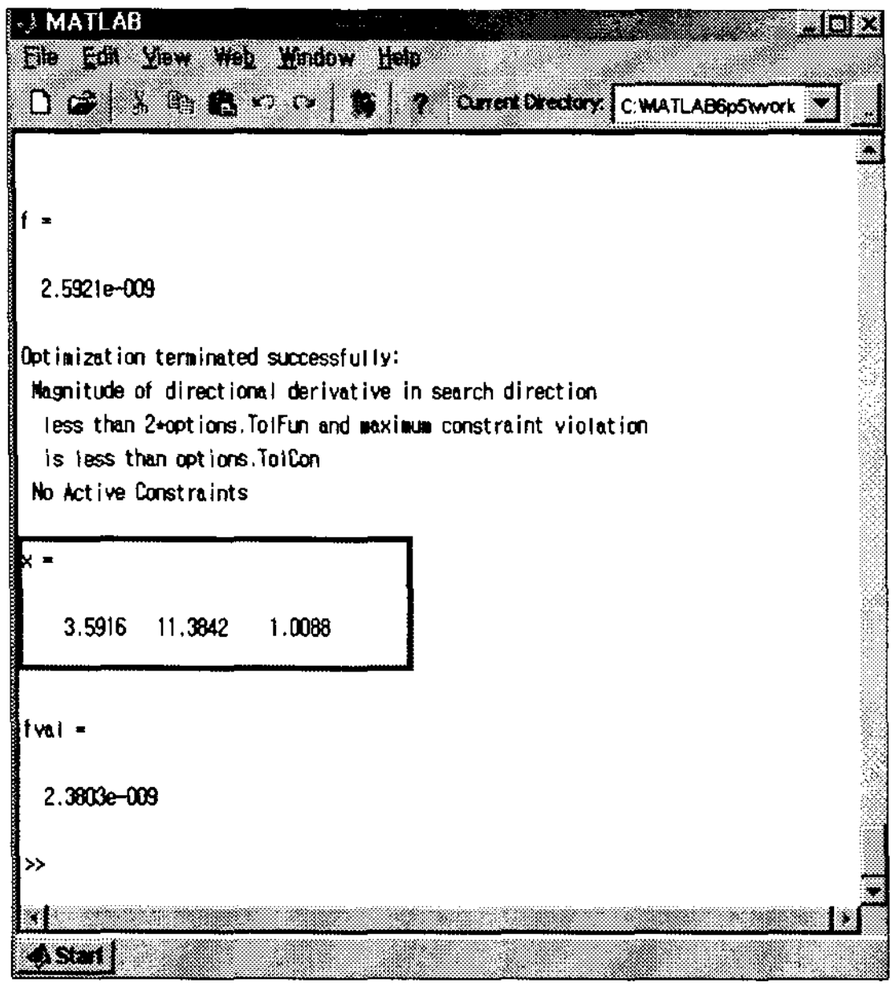
여기서 x_1= Mg, x_2= Na, x_3= Al을 나타낸다. 분산분석 결과 추정된 반응치의 회귀 식이 유의하다는 것을 알 수 있다. 이 추정식을 이용하여 목표값 τ가 1일 때, 공정의 최적조건을 다음과 같다.

$$\begin{aligned} &\text{Minimize} && [\hat{y}(x) - 1.00]^2 \\ &\text{Subject to} && x \in \Omega \end{aligned}$$

공정의 최적 조건은 <그림 4>에서 보여주는 것과 같이 MATLAB 프로그램을 사용하여 구하였다. 그 결과 최적 조건은 ($x_1 = 3.5916$, $x_2 = 11.3842$, $x_3 = 1.0088$)으로 나타났다. 여기서 fval 값은 $[\hat{y}(x) - \tau]^2$의 값을 의미한다.

표 2는 반응치 y에 대한 목표값 τ의 1부터 7

까지의 최적조건을 구한 것이다. 결과들을 살펴보면, 플로트 법으로 가공된(float processed) 유리들은($\tau=1, \tau=3$) 그렇지 않은 유리들에($\tau=2, \tau=4$) 비해 마그네슘(x_1)의 양이 많음을 알 수 있다.



<그림 4> $\tau=1$일 때, 공정의 최적조건 결과

<표 2> τ에 따른 공정의 최적조건

τ	x_1	x_2	x_3
1	3.5916	11.3842	1.0088
2	0.5131	13.7912	0.0590
3	1.4096	11.3971	1.2463
4	1.0743	12.1585	2.2932
5	0.5051	8.2844	1.0802
6	0.5031	7.6465	0.3181
7	0.5001	7.1109	0.3110

5. 결론

수많은 인자와 데이터가 발생하는 제조 현장에는 기존의 통계적 기법으로는 해결할 수 없는 많은 문제들이 발생한다. 본 연구에서는 이러한 문제를 해결하기 위하여 데이터 마이닝 기법과 통계적 기법들을 통합하여 보다 현실적이고 새로운 통계적 설계 기법을 개발하였다. 기존에 연구되어진 데이터 마이닝 기법처럼 인자간의 관련성을 규명하는데 그치는 것이 아니라 데이터 마이닝 알고리즘(CBFS와 BFS)은 통해 상관이 높은 인자를 선택하고, 이 인자들이 통계적으로 유의한가를 정확하게 알아본 다음, 수리적인 모형을 세워서 최적 조건을 구함으로써 보다 단계 발전된 기법을 개발하였다.

본 논문에서 보여주고 있는 수치예제는 단적인 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. 또한 이 기법은 순수 데이터 마이닝 학문과 응용 분야가 한층 더 발전할 수 있는 계기가 될 수 있을 것이라 믿는다.

참 고 문 헌

[1] Alexander Hinneburg, Daniel A.K.(1999), "Clustering Method for Large Data Sets". SIGMOD99

[2] Besharati, B. and Luo, L. and Azarm, S.(2006), "Multi-Objective Single Product Robust Optimization: An Integrated Design and Marketing Approach", *Journal of Mechanical Design,* Vol. 128, No.4, pp. 884-892.

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

[4] 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.

[5] 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.

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

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

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

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

[11] MATLAB: http://www.mathwork.com

[12] MINITAB: http://www.minitab.com

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

[14] Seifert, J.W.(2004), Data Mining: An

Overview. CRS Report RL31798
[15] Shin, S.M. and Cho, B.R.(2005), "Bias-specified robust design optimization and its analytical solutions", *Computer & Industrial Engineering,* Vol. 48, pp. 129-140.
[16] U. Fayyad, G. piatetsky-Shapiro, P. Smyth.(1996), "The KDD Process for Extracting Useful Knowledge from Volumes of Data", *Communication of the ACM,* Vol. 39, No. 11, pp. 27-34.
[17] Witten, I.W.H. and Frank, E. : *Data Mining: Practical Machines Learning Tools and Techniques.* 2nd edn Morgan Kaufmann, San Francisco
[18] Weka: http://www.weka.net/
[19] 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