6 시그마 위한 대용량 공정데이터 분석에 관한 연구
A Study on Analysis of Superlarge Manufacturing Process Data for Six Sigma

박 재홍, 변 재현
경상대학교 산업시스템공학부, 공학연구원

Abstract

Advances in computer and sensor technology have made it possible to obtain superlarge manufacturing process data in real time, letting us to extract meaningful information from these superlarge data sets. We propose a systematic data analysis procedure which field engineers can apply easily to manufacture quality products. The procedure consists of data cleaning and data analysis stages. Data cleaning stage is to construct a database suitable for statistical analysis from the original superlarge manufacturing process data. In the data analysis stage, we suggest a graphical easy-to-implement approach to extract practical information from the cleaned database. This study will help manufacturing companies to achieve six sigma quality.

1. 서론

최근 컴퓨터 및 센서 기술이 발달함에 따라 제조공정에서는 대용량의 제품 및 공정데이터가 실시간으로 수집되고 있다. 여기에는 제품 제조의 전과정에 걸친 공정변수의 값뿐만 아니라 최종제품의 품질변수에 이르기까지 거의 모든 데이터가 포함되어 있다. 의미 있는 분석을 하기 위한 조건은 우선 데이터를 얻는 측정시스템이 제품이나 공정을 정확히 측정할 수 있어야 하기 때문에, 측정시스템에 대한 평가가 반드시 이루어진 후에 데이터가 수집되어야 한다.([4]) 최근 제조현장에서는 측정시스템으로부터 수집된 공정데이터가 대용량인 경우, 초기(원시) 공정데이터로부터 통계적 기법을 적용하여 원하는 정보를 얻기란 쉽지가 않다. 그 이유는 결측치를 포함하고 있는 데이터가 많고, 제조공정의 불확실성과 잡음의 개입으로 인해 얻어진 데이터의 질이 낮아질 우려가 있으며, 무엇보다도 데이터의 양이 너무 방대하기 때문이다.([1]) 따라서 초기 대용량 공정데이터에 통계적 기법을 적용하기 전에 우선 데이터 손질을 한 후에 제조공정을 거쳐 생산되는 제품의 품질을 높이기 위한 분석을 수행하여야 한다.([5])

본 연구의 목적은 최근 제조공정으로부터 수집되는 대용량 공정데이터를 통계적 분석에 적합한 데이터베이스로 구축하기 위한 데이터 손질과 손질된 데이터를 기반으로 대용량 공정데이터의 특성을 고려한 체계적 분석방법을 제시함으로써 최근 국내 기업에서 활발하게 추진하고 있는 6 시그마 혁신활동의 데이터 분석방법에 도움을 주는 것이다.

2. 데이터 손질(Data Cleaning)

측정시스템으로부터 수집된 데이터가 대용량인 경우, 본격적인 통계분석을 하기 전 데이터 포맷의 정돈, 결측치의 처리, 이상치(outlier) 제거 등 통계적 분석에 용이한 데이터로 전환시키는 것이 필요하다. 이와 관련하여 Banks와 Parmigiani[5]는 대용량 공정데이터에 대한 데이터 손질을 위한 12단계의 절차를 제안하였다. 본 논문에서는 Banks와 Parmigiani의 단계를 축소하여 10단계의 데이터 손질단계를 소개함으로써 대용량 공정데이터로부터 의미 있는 정보를 얻는데 도움을 주고자 한다.

- 단계1: 데이터 입력(data input) 확인
  데이터 입력은 일관성의 유지가 중요하다. 수치데이터의 경우는 실수형(floating point)으로 문자 데이터의 경우 그대로 입력처리를 한다.
- 단계2: 시간척도에 따른 데이터 수집 확인
  데이터 기록의 식별기준인 시간에 따라 각 변수의 데이터가 수집되었는지 확인하는 단계이다.

- 단계3: 결측치(missing value) 표시
  결측치가 발생하면, 결측치를 발생원인별로 특정 코드값을 부여한다.(단계7 참조)

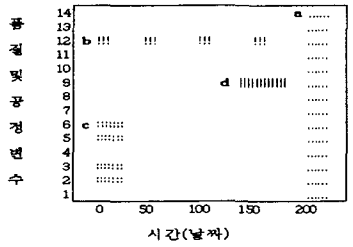
- 단계4: 데이터 크기 검사(sample size check)
  각 변수에 대한 데이터 개수를 확인하는 단계이다. 입력(input)은 예상되는 시계열의 길이와 공정데이터의 개수가 되며, 출력(output)은 각 변수들의 실제 데이터 개수와 예상 데이터 개수에 차이가 나는 변수들의 목록이다.

- 단계5: 불가능한 값(impossible value) 처리
  불가능한 값이란 일반적으로 발생할 수 없는 값을 말하며, 이상치와는 다르다. 이러한 값들에 대해서는 결측치의 처리와 유사하게 하거나, 부호(+, -)가 바뀌지 않았는지, '0'이 추가되거나 빠지지 않았는지, 데이터 입력 시 입력위치가 잘못되지 않았는지를 확인한다. 불가능한 값의 여부는 반드시 공정 전문가와 협의하여 결정하여야 한다.

- 단계6: 동기화(synchronization)
  데이터가 동기화 되도록 색인(index)을 다시 하는 단계이다. 제품의 공정변수가 품질변수에 영향을 주는 시기를 재조정하여 적합한 분석이 되

도록 한다. 예를 들어, 공정변수 값이 바뀌게 될 때 그러한 변화가 10시간 이후에 품질변수에 영향을 줄 때는 공정변수의 변화시기를 10시간 이후로 조정해야 한다.

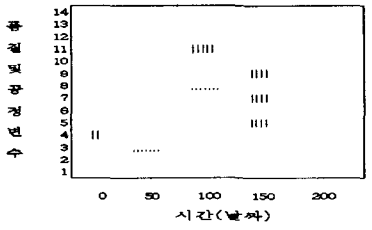
단계7: 결측치 도표 작성
결측치 도표란 각 변수들의 결측치를 시간(날짜)에 따라 작성한 것이다. 결측치의 성향을 기호로 표시하면 유용하다. 예를 들어, <그림 1>에서 a: 계획적인 공장의 조업중단, b:알 수 없는 원인 c:하위시스템의 고장 d: 돌발적인 상황이 발생했는데 발생원인을 알고 있는 경우이다. 공장 매니저는 결측치 패턴을 분석하여 공정 모니터링 시스템 개선에 유용하게 쓸 수 있다.



<그림 1> 결측치 도표 작성예제

단계8: 관측도수 차이와 결측치 추정 입력
생산공정에서 모든 공정변수가 동일한 빈도로 측정되지 않거나, 센서의 고장 또는 다른 이유로 인하여 일련의 데이터에서 간단한 gap이 생긴다. 이러한 gap을 통계적 방법과 전문가 지식 등을 이용하여 채우는 단계이다. 결측치를 대체하는 방법으로는 평균, 중앙값, 군집평균, 회귀분석 등에 의한 데이터 추정방법이 있다.([8]) Banks와 Parmigiani는 선형 보간법(linear interpolation)에 의한 데이터 추정방법을 권장한다. 그 이유는 프로그램밍 하기가 쉽고 대부분의 소프트웨어에서 수행이 빠르기 때문이다.([5])

단계9: 극한치 도표 작성
극한치 도표는 공정변수의 데이터 중 $\pm 3\sigma$를 벗어난 데이터를 도표에 표시하여 이상치를 탐지하는 데에 도움을 주는 도표이다. <그림 2>에 극한치 도표의 작성 예를 나타내었는데, 기호[|]는 $+3\sigma$을 벗어난 데이터, 기호[.]는 -3σ을 벗어난 데이터를 표시한 것이다. 발견된 이상치는 그 원인을 파악한 후 처리하고 재발하지 않도록 조치를 취해야 한다.



<그림 2> 극한치 도표 예

· 단계10. 기술통계량과 기초 탐색
시계열 데이터의 안정성, 공장중단이나 모니터링 장치의 고장이 발생한 전후의 데이터들이 일관성이 있는지를 판단하기 위하여 기초적 시계열분석, Q-Q plot 등의 도표분석을 한다. 분석대상인 특정 공정의 데이터로부터 의미 있는 정보를 추출하기 위하여 어떠한 분석 기법을 쓸 것인지를 최종적으로 점검을 하는 단계이다.

3. 대용량 공정데이터 분석

데이터 손질단계를 거친 데이터로부터 유용한 정보를 얻기 위해서는 우선 대용량 공정 데이터의 보편적인 특징을 알아야 한다. 대용량 공정데이터의 특징은 크게 4가지로 볼 수 있다. 첫째, 다단계(multistage)의 공정을 거치면서 수집된 데이터이다. 둘째, 다량의 공정변수가 관계되어 있다. 셋째, 품질변수간, 공정변수간, 품질변수 및 공정변수간 상관관계가 존재할 가능성이 크다. 넷째, 체계적인 조업실험 또는 공정제어를 통하여 얻을 수도 있지만 공정변수들이 일정범위 내에서 변동할 때 수동적으로 얻을 수 있는 데이터이다.

제조공정에서 최종제품이 생산되기까지는 여러 단계의 공정을 거치므로, 최종제품의 품질특성 변동은 모든 공정단계의 변동으로부터 영향을 받는다. 따라서 최종제품의 품질향상을 위해서는 최종단계 뿐만 아니라, 이전 단계의 공정변수들 중 중요한 변수들을 파악하여 적절한 제어를 해야만 품질특성의 변동을 감소시킬 수 있다. 일반적으로 대용량 공정데이터에는 많은 품질변수 및 공정변수가 포함되어 있으므로 일부 변수들은 크게 연관되어 있을 가능성이 크다. 따라서 공정데이터 분석의 주요 목적 중 하나는 품질변수에 크게 영향을 미치는 공정변수를 찾아내어 품질향상을 위하여 중요한 공정변수가 가져야 할 값 또는 제어범위를 결정하는 것이다. 본 논문에서는 현장 엔지니어들이 쉽게 이해하고 이용할 수 있는 공정데이터 분석방법인 covariation chart, sliced inverse box plot(SIB) chart, brushing scatter plot, box plot을 이용한 도식적인 기법을 제시하고자 한다.

3.1 Covariation Chart

covariation chart는 Banks와 Parmigiani[5]에 의해 개발된 도표로 주어진 공정데이터에 대한 품질변수간의 상관관계, 공정변수간의 상관관계, 품질변수와 공정변수 간 상관관계를 제시한다. 이 도표의 특성은 품질변수 값을 상/하위 10%로 나누어, 품질변수 값의 특성에 따른 공정변수와의 상관관계를 볼 수 있기 때문에 의미 있는 변수를 선택하는 데에 도움이 된다.

가. Covariation Chart 작성방법
(1) 품질변수와 공정변수의 선택
(2) 4 가지의 covariation chart를 작성하여 변수간 상관관계(상관계수)에 따라 백분율로 분류하여 기호로 표시[4: 90%이상, 3: 20%이상~90%미만, 2: 0.5%이상~20%미만, 1: 0.1%이상~0.5%미만, 0: 0.1%미만]([5])

위의 상관계수 분류기준은 데이터 수에 따라

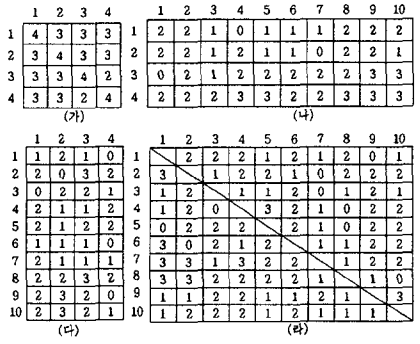
달라질 수 있다. t-검정을 이용한 상관계수의 유의성검정을 위하여 유의수준 α=0.01로써 귀무가설(ρ=0)을 기각할 수 있는 최저 상관계수를 <표 1>에 나타내었다.

<표 1> 데이터 개수의 따른 최저 상관계수

데이터 수	최저 상관계수	데이터 수	최저 상관계수
50	0.3	2000	0.06
100	0.2	3000	0.05
200	0.17	4000~6000	0.04
300	0.14	7000~10000	0.03
400	0.12	20000~50000	0.02
500	0.1	60000	0.01
600~800	0.09	70000~80000	0.009
900~1000	0.08	90000~100000	0.008

나. Covariation Chart 작성 예제

극소탄소강 철강제품의 경우를 예를 들어 covariation chart를 작성하는 방법을 예시하고자 한다.([3]) 4개의 품질변수와 10개의 공정변수를 이용하여 냉연공장 전체를 대상으로 한 5755개의 데이터를 이용하여 <그림 3>과 같은 covariation chart를 작성하였다. 분류기준은 <표 1>의 최저 상관계수와 상관계수 값의 상대적인 크기에 따라 분류하였다.[4: 90%이상, 3: 20%이상~90%미만, 2: 10%이상~ 20%미만, 1: 4%이상~10%미만, 0: 4%미만] <그림 3>에서 (가)는 품질변수들 간 상관관계를 나타내고, (나)는 품질변수들의 데이터 중 하위 10%에 해당하는 데이터에 대하여 품질변수와 공정변수 간 상관관계를 나타낸 것이며, 상위 10% 데이터에 대한 분석은 (다)에 표시되어 있다. (라)는 중요한 품질변수인 YP(항복강도)에 대하여 대각선 위쪽은 품질변수들의 데이터 중 하위 10%, 대각선 아래쪽은 상위 10% 데이터에 해당되는 부분의 공정변수간 상관관계를 나타낸 것이다.



<그림 3> covariation chart 작성 예제

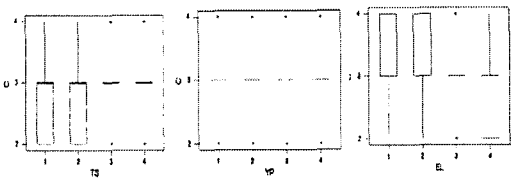
3.2 Sliced Inverse Box Plot (SIB) Chart

SIB chart는 X축에 공정변수를 놓고 Y축에 품질변수를 두는 box plot과는 달리 Y축에 공정변수 값을 놓고, X축에는 품질변수 값으로 둔다. 이렇게 변수들을 두는 목적은 품질변수 값의 변화에 따른 공정변수 분포의 변화를 보기 위함이다.([5]) 일반적으로 제품의 생산공정에는 다수의 품질변수들이 있고 각 품질변수가 바람직한 범위의 값을 갖는 제품을 생산하기 위해서는 관계되는 중요한 공정변수를 최적의 값(또는 구간)으로 제어해야 한다. 그런데 품질변수가 여러 개 있을 때에는 어떤 공정변수의 값을 결정하는 것이 쉽지 않다. 해당 공정변수의 바람직한 값이 품질변수 별로 달라지는 소위 상충(conflict)이 생기기 때문이다.

박 재홍 등[3]에서 다룬 사례에서 실제 품질변수 3개와 공정변수 3개를 대상으로 SIB chart를 이용, 각 품질변수가 좋게 나타나는 상위 %별로 구간을 나누고, 각 구간별 공정변수별 분포를 분석한다. 품질변수는 TS(27.5 kg/mm^{2}이상), YP(17.5 kg/mm^{2}이하), EL(37%이상), 공정변수는 C(탄소), S(황), TI(티타늄)을 분석 대상으로 하였다. <표 2>는 각 품질변수별로 바람직한 값의 방향을 기준으로 하여 품질변수의 값을 구간으로 나타낸 것이다. <그림 4>는 각 품질변수의 구간별로 C의 분포를 나타낸 것이다. 이러한 방법으로 2개의 다른 품질변수에 대해서도 SIB chart 분석을 할 수 있고, 최종적으로 공정변수의 값 또는 제어범위는 각 품질변수가 규격을 벗어날 때의 손실비용을 고려하여 결정해야 하겠다.

<표 2> 3개의 품질변수별 구간설정

구간번호	spec(%)	TS	YP	EL
1	90%이상	27~27.6	17~17.5	44~45.1
2	70%~90%	27.7~28.8	16~16.9	45.2~47.3
3	50%~70%	28.9~30.5	15~15.9	47.4~49.5
4	50%이하	30.1이상	14.9이하	49.6이상



<그림 4> 품질변수의 구간별 C의 분포

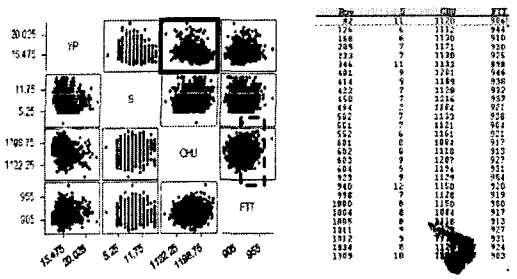
3.3 Brushing Scatter Plot

brushing이란 다차원 데이터(multidimensional data) 일부의 변수간 상관관계를 분석하기 위한 방법으로, 어떤 특정 영역의 데이터에 대하여 변수들 간 상관성을 scatter plot matrix 상에 그래프로 나타내는 방법이다.([6]) brushing에는 여러 가지 방법이 있으나, 본 절에서는 Minitab에서 적용 가능한 highlight의 transient mode에 대해 설명하고자 한다. highlight의 transient mode란 분석하고자 하는 측정점이나 측정점들의 영역을 지정하고(highlight), 그 영역을 이동해 가며(transient), 각 영역별 변수간 상관관계를 그래프로 확인하는 방법이다.

가. 이상치(outlier) 특성 파악

<그림 5>는 박 재홍 등이 다룬 사례[3]에서 1냉연공장 1522개의 데이터를 대상으로 YP(항복강도)와 CHU(추출온도)에 관한 2차원 산점도에서 이상치를 highlight하여 이러한 이상치를 발생하게 하는 주요 원인이 되는 공정변수를 파악하는 그림이

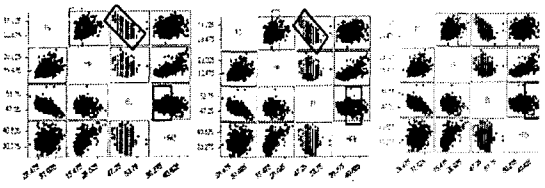
다. 이 그림으로부터 CHU외에 FFT(열연온도, TOP부)가 주요원인임을 알 수 있다. 오른쪽 표를 보면 YP의 이상치를 발생시키는 FTT의 범위는 높은 구간(920~950 °C)임을 알 수 있다.



<그림 5> 이상치의 원인파악을 위한 예

나. highlight의 transient mode 적용

scatter plot matrix에서 자신이 분석하고자 하는 속성에 따라 데이터를 구간으로 분류하고, 각 구간에 있는 데이터를 highlight하여 transient mode를 적용하면 각 구간별로 공정변수의 값을 어떻게 가지고 가야 하는지를 알 수 있다. <그림 6>에서 점선으로 표시된 부분은 highlight 된 영역이고 실선은 분석을 위해 표시한 부분이다. 그림을 보면 YP가 하한구간이나 중간구간에 있을 때, TS와 EL이 음의 상관관계가 있음을 알 수 있다.



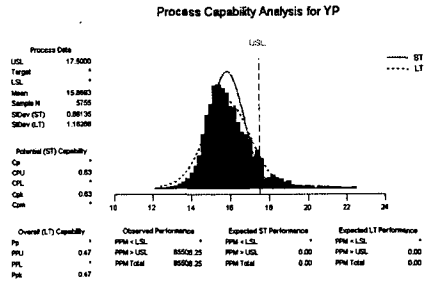
<그림 6> 구간별 transient mode 예

3.4 Box Plot

대용량 공정데이터로부터 공정능력을 파악하여 시그마 수준으로 환산해 봄으로써 현재 원하는 품질수준을 가진 제품을 어느 정도 생산하고 있는지 알 수 있다. 본 절은 box plot을 이용하여 최적 공정조건을 찾는 방법을 사례를 통해 소개하고자 한다. 본 사례는 극저탄소강의 재질 중 개선이 필요한 YP(항복강도)를 품질특성으로 선정하여 그 수준을 하향 안정화시키기 위해 주요 공정변수의 제어범위를 정하는 것이다.([3])

가. 전체공정의 공정능력 분석

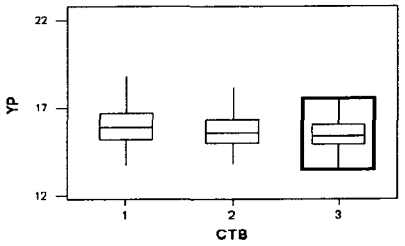
<그림 7>은 박 재홍 등[3]이 사례에서 다룬 전체냉연공정의 YP에 대한 공정능력분석 결과이다. 분석결과를 보면 Cpk = 0.63이고, Ppk = 0.47로서, 시그마수준은 2.91 σ 수준[(Ppk×3)+1.5]이 된다.



<그림 7> YP의 공정능력분석

나. 최적구간의 도출

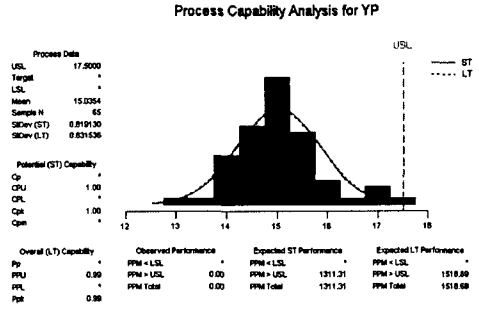
공정변수의 제어범위를 3개의 구간(1: 하한구간, 2:중간구간, 3: 상한구간)으로 분류하고 구간별로 품질변수의 box plot을 도시하여 YP의 값을 가장 크게 줄일 수 있는 구간을 최적구간으로 선정한다. <그림 9>에 냉연공정의 공정변수인 권취온도(CTB)를 예로 들었다. YP를 적정 제어범위인 17.5kg/mm^{2}이하로 두기 위해서는 권취온도를 '상한구간'으로 제어해야 됨을 알 수 있다. 다른 주요 공정변수들의 최적구간도 같은 방식으로 구하여, 전체적인 최적 공정조건을 구할 수 있다.



<그림 8> CTB의 최적구간 도출예제

다. 최적구간 검증

최적구간의 타당성 검증을 위하여 전체 데이터 중에서 각 주요 공정변수들의 최적구간으로 구성된 데이터를 추출하여 YP에 대한 공정능력을 분석하여, 현 공정에 대한 시그마 수준과 최적구간에서 시그마 수준을 평가해 보았다. 그 결과 시그마 수준이 2.91 σ 에서 4.47 σ 로 향상되었다.



<그림 9> 최적구간의 YP 공정능력

4. 결론 및 추후 연구과제

본 논문에서는 제조공정에서 발생하는 대용량 공정데이터를 이용하여 의미 있는 품질정보를 추출하기 위하여 필요한 절차를 제시하였다. 우선 통계적 분석에 적합한 데이터베이스를 구축하기 위한 데이터 손질 방법을 Banks와 Parmigiani[3]에 기초하여 제안하였다. 그리고 현장의 엔지니어들이 쉽게 이용할 수 있는 도식적인 방법을 제시함으로써 대용량의 데이터를 다루는 국내 기업이 6시그마 품질수준에 접근할 수 있도록 하였다. 대용량 공정데이터는 제조공정별로 그 특성이 다를 수 있으므로 분석대상에 따라 본 논문에서 제시하는 방법을 수정하여 사용하거나 적합한 분석방법을 추가하여 이용할 수 있으리라고 본다.

참고문헌

[1] 김 영상(1999), "공정모니터링 데이터 분석을 위한 편차최소제곱법과 인공신경망의 비교 연구", 한국과학기술원 산업공학과 석사학위논문.

[2] 박 성현(1998), *회귀분석*, 민영사.

[3] 박 재홍, 변 재현, 김 창현, 정 창원, 최 영대(2001), "구간세분화 방법을 이용한 철강산업체의 6시그마 프로젝트 추진사례", *품질혁신*, 제2권, 제1호, pp.57-66.

[4] 배 도선 외 6인(1999), *통계적 품질관리*, 영지문화사.

[5] Banks, L. D., Parmigiani, G.(1992), "Pre-Analysis of Superlarge Industrial Data Sets", *Journal of Quality Technology*, Vol.24, pp.115-129.

[6] Becker, R. A., Cleveland, W. S.(1987), "Brushing scatterplots", *Technometrics*, Vol.29, pp. 115-129.

[7] MINITAB(2000), *Minitab Statistical Software: User's Guide 1-2*, MINITAB Inc., release 13.

[8] Pyle, D.(1999), *Data Preparation for Data Mining*, Morgan Kaufmann Publishers.