잠재변수에 대한 규칙추론을 통한 공정 최적화

Process optimization using a rule induction method based on latent variables

정일교[1], 이상호[2], 전치혁[2*]

[1]삼성전자 반도체연구소 포토마스크팀, [2]포항공과대학교 산업경영공학과

E-mail: ilgyo.chong@samsung.com, samo35@postech.ac.kr, chjun@postech.ac.kr

Abstract

In order to determine new settings of key process variables optimally, a new rule induction method through a historical data is proposed without using an explicit functional model between process and quality variables. First, a partial least square is used to reduce the dimensionality of the process variables. Then new process settings that yield the best quality variable are identified by sequentially partitioning the reduced latent variable space using a patient rule induction method. The proposed method is illustrated with a case study obtained from steel-making processes. We also show, through simulation, that the proposed method gives more stable results than estimating an explicit function even when the form of the function is known in advance.

Key words: data mining; process optimization

1. 서론

제품의 품질을 높이기 위해서 적정 공정조건을 찾는 문제는 엔지니어에게 매우 중요하다. 이러한 목적으로 두 가지 방법이 있을 수 있는데, (1) 함수적 접근 방법과 (2) 비함수적 접근 방법이다. 함수적 접근 방법은 우선 품질변수와 공정변수 간의 품질 모델을 가정하고 데이터로부터 모델 파라미터를 추정하고 최적화함으로써 적정조건을 찾는다. 비함수적 방법은 품질변수와 공정변수 간의 품질 모델 추정 없이 곧 바로 다량의 데이터로부터 공정인자의 적정조건을 추론한다.

제조 공정은 변수간의 상관관계가 강하고, 품질인자와 공정인자간의 관계가 복잡하다. 이러한 경우에 함수적 접근 방법은 품질인자와 공정인자 간의 관계를 추론한 모델이 실제 관계와 많이 다를 수 있으며, 따라서 추론된 모델로부터 구해진 해를 신뢰하기 힘들다. 반면에 비함수적 접근 방법은 해를 공정 데이터 안에서 찾기 때문에 데이터에 덜 민감할 수 있다.

본 논문은 공정인자의 적정 조건을 찾기 위한 새로운 비함수적 접근방법을 제안한다. 2장에서는 제안기법의 개념과 알고리즘을 설명하고, 3장에서는 실제 적용 사례를 소개하고, 4장에서는 제안된 기법을 함수적 방법과 비교하기 위한 가상실험을 수행한다. 마지막으로 5장은 본 연구에 대한 결론을 내린다.

2. 잠재변수를 이용한 규칙추론 방법

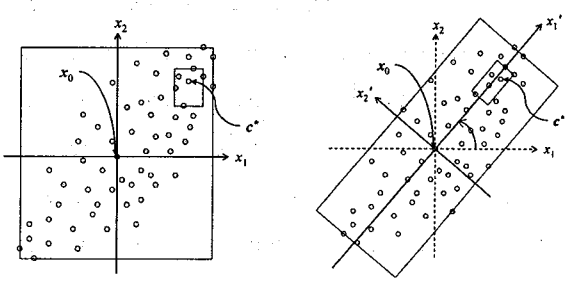
2.1 제안 기법의 기본 개념

인내심 규칙추론 방법(Patient Rule Induction Method; PRIM)은 함수적인 접근 방법의 문제점을 극복하기 위한 대안으로 Friedman과 Fisher(1999)가 제안했으며, 금융, 생물정보학, 마케팅 등 다양한 분야의 성공 적용 사례가 있다. PRIM은 [그림1-a]에서와 같이 모든 데이터를 포함하는 큰 상자를 정의한 후, 상자를 여러 번 깎아서 작은 상자를 최종 결과로 얻는 방법이다. [그림1-a]에서 작은 상자는 PRIM에 의해서 큰 상자로부터 얻어진 것을 나타낸다. 공정 최적화 문제에서는 이러한 상자들로부터 공정인자의 적정수준을 찾을 수 있다.

상자를 깎을 때는 상자의 면에 수직이 되는 모든 방향을 고려하며, 이러한 방향대로 깎았을 때에 보다 좋은 특성(상자의 목적 값으로 표현되고 2.2에서 언급)의 데이터들을 포함하는 방향으로 상자를 깎는다. 예를 들어 [그림1-a]에서와 같이 변수(공정인자)가 2개가 있다면, 상자는 2차원에서 정의가 되고 각 면의 왼쪽을 깎는 경우와 오른쪽을 깎는 경우가 존재하므로 총 4(=2×2)개의 깎을 수 있는 방향이 있고, 이러한 4개의 방향대로 깎았을 때에 상자에 남아 있는 데이터들의 특성이 제일 좋은 방향으로 상자를 깎는다. 이와 같은 방법을 충분히 작은 상자가 얻어질 때까지 반복한다.

상자를 깎는 량은 PRIM에서 매우 중요한 역할을 하는데 현재 상자에 포함되는 데이터 수의 α% 미만이 잘려지는 수준에서 결정되며 일반적으로 α는 0.05에서 0.1사이의 숫자를 사용한다. α의 값으로 작은 숫자를 사용하면 큰 상자로부터 작은 상자를 얻기 위해 상자를 깎아야 하는 횟수가 증가한다. 상자를 깎는 횟수가 많다는 것은 한번의 깎는 결과가 최종 결과에 큰 영향을 주지 않는 다는 것을 의미하고 이는 PRIM이 다른 알고리즘에 비해 안정적인 결과를 도출하는 이유기도 하다. 한편, α로 작은 값을 사용하는 것은 알고리즘 소요시간이 오래 걸리기 때문에 '인내심 전략' 이라고 불린다.

PRIM은 내부적으로 변수가 독립이라 가정한다. 그런데 공정인자는 대개의 경우에 서로 상관관계가 높기 때문에 PRIM의 가정에 위배된다. 따라서 본 논문에서는 [그림1-b]에서와 같이 PCA나 PLS를 이용하여 서로 독립이 되는 새로운 변수를 만들고(Geladi and Kowalski, 1986), 새로운 변수로 정의되는 공간에 PRIM을 적용하여 공정인자의 적정 조건을 찾는다.



(a) PRIM (b) 제안된 기법

[그림1] PRIM과 제안된 기법의 개념 비교

2.2 알고리즘

본 섹션에서는 제안된 방법을 설명하기 위한 용어와 알고리즘을 간략히 설명한다. 제안된 방법에 대한 자세한 알고리즘은 정일교(2006)를 참고한다.

● 상자에 대한 평가지수

한 개의 품질인자(y)와 p개의 공정인자($\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$)에 대한 N개의 관측치로 이루어진 공정 데이터 $\{(y_i, \boldsymbol{x}_i), i = 1, 2, \ldots, N\}$와 임의의 상자 B가 주어졌을 때 상자의 지지도, 목적 값 그리고 공정지수는 다음과 같이 구해진다.

(1) 상자의 지지도 (β_B)

$$\beta_B = (1/N)\sum_{i=1}^{N} 1(\boldsymbol{x}_i \in B) \quad (1)$$

상자의 지지도는 전체 N개의 데이터 중에서 상자B에 포함되는 데이터의 비율을 나타낸다.

(2) 상자의 목적 값 (Obj_B)

$$Obj_B = (1/n_B)\sum_{x_i \in B} h(y_i) \quad (2)$$

상자의 목적 값은 식 (2)와 같이 상자 B에 포함되는 y 데이터에 함수$h(\cdot)$를 취한 평균으로 표현된다. n_B는 상자 B에 포함되는 데이터의 개수이고, $h(\cdot)$는 문제의 상황에 따라 정의하며 우리가 찾는 데이터를 많이 포함할수록 큰 값을 갖도록 정의한다.

(3) 상자의 공정지수 (F_B)

임의의 상자 B로부터 공정최적조건 $\boldsymbol{x}^*$을 추정했다고 하자. 이 때 상자의 공정지수는 우선, 충분히 많은 R개의 데이터를 공정최적조건 $\boldsymbol{x}^*$를 중심으로 무작위로 발생시키고 그 중에서 상자 B에 얼마나 포함되는지를 식(3)과 같이 계산한다.

$$F_B = \sum_{i=1}^{R} 1(\boldsymbol{x}_i \in B)/R \quad (3)$$

데이터를 무작위로 발생시킬 때는 공정인자들 간의 상관관계를 과거데이터로부터 추정할 수 있다. 상자의 공정지수(F_B)는 상자의 지지도와 달리 미래에 공정을 $\boldsymbol{x}^*$로 바꿨을 때 현장에서 상자 B안에서 공정을 제어할 수 있는 능력을 지수화한 것이다.

● 슈도 코드

단계1. 공정데이터를 러닝과 테스트 셋으로 나눈다.

단계2. 테스트 셋을 이용하여 공정데이터에 PLS 또는 PCA를 적용하여 잠재변수를 생성하고 잠재변수의 수를 결정한다. 일반적으로 잠재변수의 수는 공정인자의 수 보다 적기 때문에 차원 축소의 효과가 있다.

단계3. 잠재 데이터에 대해서 PRIM을 적용하여 큰 상자에서 작은 상자가 얻어 질 때까지의 상자 열($B_1, B_2, \cdots, B_K$)을 얻는다. 여기서 $B_1 \supset B_2 \supset \cdots \supset B_K$ 임을 기억하자.

단계4. 각 상자로부터 공정인자의 적정조건($\boldsymbol{x}_k$, k=1,2,…,K)을 계산하여 (($B_1, \boldsymbol{x}_1$), ($B_2, \boldsymbol{x}_2$), …, ($B_K, \boldsymbol{x}_K$))을 얻는다. 본 논문에서는 상자의 중심으로 적정조건을 계산한다.

단계5. 테스트 셋을 이용하여 (($B_1, \boldsymbol{x}_1$), ($B_2, \boldsymbol{x}_2$), …, ($B_K, \boldsymbol{x}_K$))에 대한 상자의 지지도, 상자의 목적 값, 그리고 상자의 공정지수를 계산하고 가장 좋은 특성을 갖는 상자와 적정조건을 선택한다.

3. 사례연구

본 사례 연구의 목적은 한 개의 품질인자가 가능한 작은 값을 갖게 하기 위해 7개의 공정인자의 적정 조업조건을 결정하는 것이다. 제안된 기법을 적용하여 총 33개의 상자와 해당 적정 조업조건 (($B_1, \boldsymbol{x}_1$), ($B_2, \boldsymbol{x}_2$), …, ($B_{33}, \boldsymbol{x}_{33}$))을 얻었다.

[표1] 제안기법을 이용한 조업조건 결정

	Obj_B	β_B	F_B
B_1	-0.0553	0.997	1.0
B_2	-0.0451	0.890	1.0
B_3	-0.0376	0.801	1.0
…	…	…	…
B_{16}	-0.0124	0.194	0.944
B_{17}	***-0.0108***	***0.171***	***0.941***
B_{18}	-0.0110	0.152	0.933
…	…	…	…
B_{33}	-0.0070	0.031	0.376

본 사례에서 7개의 공정인자가 있으나 PLS를 적용하여 2개의 새로운 잠재변수에 대

해서 상자들을 만들었다. 식(2)의 함수 h는 '$h(y_i)=-(y_i-Target)^2$'을 사용했으며 Target은 1.0으로 본 사례의 품질인자가 취할 수 있는 충분히 작은 값이다. [표1]은 33개 상자에 대한 목적값, 지지도 그리고 공정지수를 요약한 것이다. 총 33개의 상자 열 중에서 전체 데이터의 10% 이상을 포함하고 공정지수가 90% 이상 되는 상자들 중에서 상자의 목적 값이 제일 큰 17번째 상자를 최종 결과로 결정하였다. B_{17}은 첫 번째, 두 번째 잠재변수에 대해 각각 $-3.685 \le t_1 \le -0.694$, $-0.932 \le t_2 \le 0.526$로 정의되고 이에 해당하는 공정인자의 적정조건은 x_{17}^*은 (0.0513, 3.13, 0.28, 0.027, ..., -43)이다. 추가 실험이 필요하겠지만 과거 데이터에 기반으로 B_{17}은 약 7.9%의 품질개선이 예상된다.

4. 가상실험

함수를 이용하는 방법, PRIM을 바로 적용하는 방법 그리고 제안된 방법의 성능 비교를 위한 가상실험을 수행하였다.

$$y_i = (\boldsymbol{x}_i - \boldsymbol{c}^*)\Gamma^{-1}(\boldsymbol{x}_i - \boldsymbol{c}^*)^t + \varepsilon_i, (i=1,2,..,N) \quad (4)$$

가상실험을 위한 데이터는 식(4)의 모델로부터 생성한다. 식(4)에서 실험요인은 (i) 공정인자간의 상관관계 Γ, (ii) 데이터의 개수 N, 그리고 (iii) 노이즈의 크기 ε이다. 가상실험에서 공정인자는 7개를 가정하였고 현재 공정인자의 셋팅값은 (0,0,0,0,0,0,0)으로 가정하였다. 또 품질인자의 *Target*은 0으로 가정하였으며 식(4)에서 보듯이 공정인자 값이 $\boldsymbol{c}^*$ 일 때 품질인자 값이 평균적으로 *Target* 값이 되도록 하였다.

함수를 이용하는 방법을 적용할 때는 실제 현장에서는 함수의 형태를 알 수 없음에도 불구하고 함수의 형태가 알려져 있다고 가정한 뒤, 함수의 계수만을 데이터로부터 추정하고 최적화 함으로써 해를 구했다. PRIM을 적용할 때는 Friedman이 제안한 방법을 공정최적화에 적합하게 적용하였으며, 제안된 기법을 적용할 때는 먼저 PLS를 적용하여 새로운 잠재변수를 구하고 잠재변수에 대해서 PRIM을 적용하여 최적해를 구하였다.

실험인자에 대해 각각 2개, 3개, 3개의 실험수준을 갖고 500번의 반복 실험을 수행하여 총 9,000번의 가상실험을 하였다. 각 공정최적화 기법을 평가할 때는 9,000번 중에 개선해를 찾은 비율(즉, 현재해 (0,0,0,0,0,0,0)보다 $\boldsymbol{c}^*$에 더 가까운 해를 찾은 비율)로써 비교하였다. [표2]에서 보듯이 제안된 기법은 대부분의 상황에서 개선해를 찾았다.

[표2] 가상실험 결과

	함수적 방법	PRIM	제안기법
개선해 비율	92.3 %	88.9 %	98.8 %

5. 결론

본 논문은 공정최적화를 위해 품질인자와 공정인자의 함수를 추정하지 않고 곧 바로 다량의 데이터로부터 데이터마이닝 기법을 이용하여 공정인자의 적정수준을 찾는 새로운 방법을 소개하였다. 추후 연구로 PCA와 PLS를 적용했을 때의 성능 비교 그리고 품질인자가 다수일 때의 연구가 필요하다.

참고문헌

[1] Friedman JH, Fisher NI. (1999) Bump hunting in high-dimensional data. *Statistics and Computing*, **9**, 123-143

[2] Geladi P, Kowalski BR. (1986) Partial Least-Sqaures Regression: A Tutorial. *Analytica Chimica Acta*, **185**, 1-17

[3] 정일교. (2006) 공정 최적화의 데이터 마이닝 접근 방법: 다중공선성을 고려한 규칙 추론. 박사학위논문, 포항공과대학교