응용통계연구 (2010)
23(2), 295–304

부분최소자승법과 변수선택을 이용한 코팅두께 예측모델 개발

이혜선[1] · 이영록[2] · 전치혁[3] · 홍재화[4]

[1]포항공과대학교 산업경영공학과, [2]포항공과대학교 산업경영공학과
[3]포항공과대학교 산업경영공학과, [4]포스코기술연구소

(2010년 1월 접수, 2010년 2월 채택)

요 약

산업체 공정과정에서 타겟품질변수의 실시간 예측과 관리는 품질제고, 수익율 향상에 중요한 관건이 된다. 본 연구는 내지문강판의 코팅두께를 비파괴적이고 신속한 방법으로 예측하여 균일한 품질의 강판을 생산하기 위해 UV스펙트럼데이터를 이용한 최적예측모델을 개발하고자 한다. 부분최소자승법에서 변수중요도척도를 이용한 변수선택방법은 노이즈성 영역의 독립변수를 줄임으로써 예측정확도는 높일 수 있으며, 스펙트럼데이터의 경우 원데이터보다 적절한 데이터전처리가 예측정확도를 높이는 정보를 제공하기도 한다. 본 연구에서는 부분최소자승법 예측모델에서 변수선택방법과 데이터전처리효과가 내지문강판 코팅두께 예측정확도 향상에 기여하는 결과를 제공하고, 스펙트럼데이터를 이용한 품질변수 예측모델 개발 시 적용할 수 있는 일반적인 변수선택방법과정을 제안한다.

주요용어: 부분최소자승법, 변수중요도척도, 데이터전처리.

1. 서론

스펙트럼데이터와 같은 고차원 독립변수들을 이용한 예측은 노이즈, 다공선성 등의 문제로 정확하고 안정적인 모델개발이 어려운 경우가 많다. 특히 복잡한 공정을 거치는 공정데이터 혹은 분광데이터를 이용한 예측모델에서 적절한 데이터전처리와 변수선택은 품질변수의 예측정확도에 매우 중요한 영향을 갖는다. 고차원 데이터를 이용한 예측모델 시 변수선택과정 혹은 변환 등은 노이즈를 감소시키고 데이터 차원을 낮추는 유용한 과정이다. 근적외선, 적외선, 엑스선회절 등의 방법에 의해 얻은 스펙트럼데이터를 이용한 예측모델은 의약, 공학, 약학 등 많은 분야에서 이루어져왔다. 스펙트럼데이터는 고차원데이터이며 특성상 일부영역에서는 노이즈를 갖고 있어서 변수선택 혹은 변수구간선택에 의해 예측정확도를 향상시킬 수 있다. 엑스선회절 데이터를 이용한 환원율예측에서 변수중요도척도를 이용한 변수선택방법은 전체 독립변수의 30%만을 사용해서도 전체변수를 사용했을 때와 유사한 오차를 갖는다는 것을 보여준 연구 (Lee 등, 2007)가 있으며, 변수 간 다중공선성이 존재할 때 부분최소자승법의 변수중요도척도를 이용한 변수선택이 예측성능이 우수하다는 시뮬레이션 연구 (Chong과 Jun, 2005)가 있으며, 오일의 순도를 예측할 때 라만스펙트럼을 사용하는데 전체구간보다는 일부 최적구간을 선택이 예측에 효율적임을 보인 연구 (Heise 등, 2005; Cramer 등, 2008)가 있다. 이 연구에서는 UV스펙트럼을 이용하여 내지문강판의 코팅두께를 예측하는 문제를 다루고자 한다. 부분최소자승법을 사용하는데 연속된 구간을 선택해야 한다는 제약조건에서 예측정확도를 향상시키는 최적변수구간선택방법을 제안하고자 한다. 또한, 변수중요도척도(variable importance in projection; VIP)를 이용한 변수선택 뿐 아니라, 데이터 전처리 효과가 예측의 정확도에 미치는 영향을 분석하여 최적분석방법을 제시하고자 한다.

[1]교신저자: (790-784) 경북 포항시 남구 효자동 산 31, 포항공과대학교 산업경영공학과, 연구교수.
E-mail: hyelee@postech.ac.kr

2. 부분최소자승법과 변수중요도척도

부분최소자승법은 X-space상에서의 분산을 설명하는 성분을 도출할 때 종속변수 Y와의 상관관계를 고려하여 가중치를 구하는 반복과정을 거쳐 잠재변수를 도출한다 (Wold 등, 2001). 이 점은 주성분분석이 X-space상의 분산을 최대로 설명하는 성분을 구하는 것과 차별된다. 데이터 차원이 높은 경우, 부분최소자승법은 주성분회귀보다 예측력이 높고, 노이즈 제거 효과도 우수하다. 변수수가 p이고 관측수가 n일 때 $\mathbf{X}$는 $(n \times p)$행렬이고 $\mathbf{Y}$는 $(n \times 1)$이면, 부분최소자승회귀모형은 다음과 같이 표현된다.

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{2.1a}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^T + \mathbf{F} \tag{2.1b}$$

$$\mathbf{u}_a = b_a \mathbf{t}_a + \mathbf{h}, \quad a = 1, \ldots, A, \tag{2.1c}$$

여기서 A는 잠재변수의 수, $\mathbf{t}_a$와 $\mathbf{u}_a$는 각각 $\mathbf{X}$와 $\mathbf{Y}$의 a번째 잠재벡터이며 $\mathbf{T}$와 $\mathbf{U}$는 스코어행렬이다. $\mathbf{P}$와 $\mathbf{Q}$는 $\mathbf{X}$와 $\mathbf{Y}$의 로딩행렬이며 $\mathbf{E}$, $\mathbf{F}$, $\mathbf{h}$는 각 모델에 해당하는 오차항을 나타내는 행렬 및 벡터이다. b_a는 $\mathbf{T}$와 $\mathbf{U}$간의 회귀계수로서 NIPALS(nonlinear iterative partial least squares)알고리즘으로 추정된다 (Wold 등, 1984). 잠재변수의 수의 결정은 예측변수의 정확도에 매우 중요하며 이는 cross validation에 의해 결정한다.

변수중요도척도(Variable Importance in Projection; VIP)는 개별독립변수 $\mathbf{x}_j$가 잠재변수 $\mathbf{t}_a$들에 기여한 정도를 나타내는 척도로써, $\mathbf{Y}$의 총변동중 $\mathbf{t}_a$에 기여한 가중치를 합산하여 산정한다.

$$\mathbf{VIP}_j = \sqrt{p \sum_{a=1}^{A} w_{ja}^2 b_a^2 \mathbf{t}_a^T \mathbf{t}_a \Big/ \sum_{a=1}^{A} b_a^2 \mathbf{t}_a^T \mathbf{t}_a}, \tag{2.2}$$

w_{ja}는 a번째 잠재변수에서 j번째 독립변수가 갖는 비중이며 VIP 제곱값의 평균이 1이므로 일반적으로 1보다 큰 VIP값을 가지면 유의한 변수후보로 본다. VIP값이 특정 기준값보다 큰 영역의 흡광도만을 독립변수로 사용하는 변수선택과정은 흡광도($\mathbf{X}$)의 측정비용 절감뿐만 아니라, 전체 파장 구간을 사용하는 경우와 비교했을 때 노이즈 제거효과로 예측 정확도를 향상시킬 수 있다.

3. 데이터전처리와 변수선정

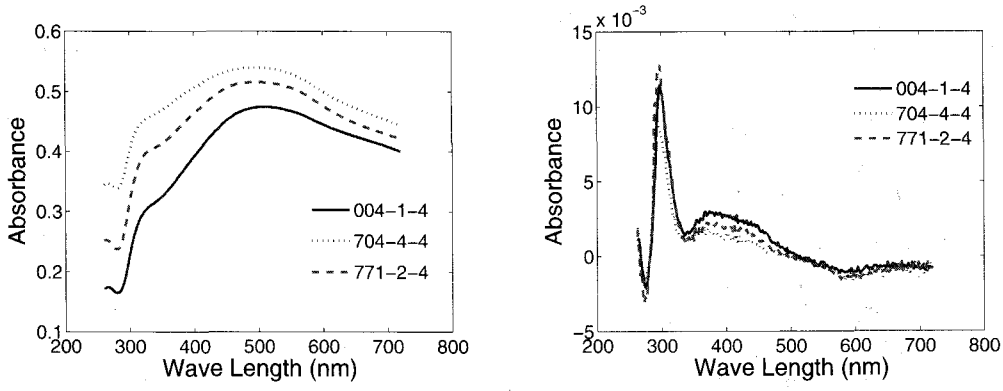
3.1. 실험 데이터

코팅두께를 비파괴적이며 신속하게 예측하기 위하여 자외선-가시광선(UV; ultraviolet-visible)영역의 스펙트럼을 이용하였다. 스펙트럼의 측정은 IMS사의 Table top형 기기를 이용하고 Xe lamp에서 발생된 UV 광원을 이용하여 시료에 optical fiber를 통하여 얻어졌다. 반사된 스펙트럼은 polychromator를 거쳐 7nm(nanometer)의 해상도를 가지는 Hamamazu사의 PDA(Photo Diode Array) 검출기를 이용하여 파장 범위 260~720nm에서 2nm 간격으로 얻은 231차원의 흡광도를 얻었다. 각 파장에서의 흡광도가 독립변수가 되고 코팅두께가 종속변수가 된다. 예측되는 코팅두께의 소재는 내지문(anti-fingerprint)강판으로써, 내지문강판은 지문이나 오염물질이 잘 묻지 않고 내식성 및 표면외관이 우수하여 컴퓨터, 음향기, 복사기 등의 Case 및 부품 소재로 널리 사용되며, 아연도금층 위에 크롬처리공정을 거친 후 내지문수지 처리를 한 강판이다. 생산과정에서 실리콘계의 유기피막코팅을 입히는데 너무 두껍거나 얇게 되는 경우 Coater의 속도와 압력을 조정하여 품질을 조절한다.

본 연구에서는 14개 내지문강판 샘플을 선정하여, 각각의 유기피막코팅 두께를 측정하였다 (표 3.1). 각 샘플의 흡광도를 4회 반복하여 측정하여 총 56개의 스펙트럼을 얻었다. 얻어진 흡광도 데이터 $\mathbf{X}$는 아

표 **3.1.** 내지문강판 샘플

샘플번호	스펙트럼 번호	스펙트럼 ID	코팅두께(mg/m^2)
1	1 ~ 4	004-1-1 ~ 004-1-4	940.0
2	5 ~ 8	004a-4-1 ~ 004a-4-4	751.5
3	9 ~ 12	004a-5-1 ~ 004a-5-4	810.3
4	13 ~ 16	007-3-1 ~ 007-3-4	1294.9
5	17 ~ 20	703-3-1 ~ 703-3-4	1058.1
6	21 ~ 24	703-4-1 ~ 703-4-4	648.7
7	25 ~ 28	704-4-1 ~ 704-4-4	605.4
8	29 ~ 32	771-2-1 ~ 771-2-4	1108.9
9	33 ~ 36	771-5-1 ~ 771-5-4	700.6
10	37 ~ 40	776-1-1 ~ 776-1-4	1222.3
11	41 ~ 44	776-2-1 ~ 776-2-4	1267.2
12	45 ~ 48	776-5-1 ~ 776-5-4	1150.1
13	49 ~ 52	1385427-1 ~ 1385427-4	861.8
14	53 ~ 56	1385519-1 ~ 1385519-4	1009.9



(a) 1차 차분 전 스펙트럼 (b) 1차 차분 후 스펙트럼

그림 **3.1.** 내지문강판의 흡광도 스펙트럼

래 식과 같이 정의된다.

$$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_{56}]^T,$$

where

$$\mathbf{x}_i = \left(x_i^{(1)}, \ldots, x_i^{(231)}\right)^T, \quad i = 1, \ldots, 56,$$

$x_i^{(j)}$: absorbance for the wavelength of $260 + 2(j-1)$nm, $\quad i = 1, \ldots, 56,\ j = 1, \ldots, 231.$

3.2. 데이터 전처리

흡광도의 변동량을 보면 자외선 영역(~380nm)에서 가시광선 영역으로 파장이 길어지는 동안 흡광도가 급격히 증가하다가 blue(450~495nm)~green(495~570nm)영역에서 최대가 되고 이후 서서히 감소하는 완만한 형태의 스펙트럼이 얻어졌다 (그림 3.1(a)).

일정간격에서 얻은 흡광도간에는 상관관계가 높으므로 다중공선성의 영향을 감소시키기 위해 1차 차분

의 데이터 전처리를 아래와 같이 수행하였다. 1차 차분은 스펙트럼 데이터를 이용한 예측모델에서 특징적인 정보를 추출하기 위하여 시도되는 전처리방법 중 하나이다. 파장축으로 일차차분을 한 이유는 시계열 자료에서 일차차분을 하는 이유와 유사하게 보다 안정적인 데이터를 만들기 위한 것이다. 2차 차분 등을 고려할 수 있으나 고차 차분의 경우 오히려 데이터의 특성정보의 손실을 가져올 수 있다.

$$\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_{56}]^T,$$

where

$$\mathbf{z}_i = (z_i^{(1)}, \ldots, z_i^{(230)})^T, \quad i = 1, \ldots, 56,$$
$$z_i^{(j)} = x_i^{(j+1)} - x_i^{(j)}, \quad i = 1, \ldots, 56, \; j = 1, \ldots, 230.$$

1차 차분된 데이터는 차분 이전에 비해 1차원이 감소한다. 그림 3.1(b)와 같이, 파장 300nm 전후로 차분값이 급격히 증가하다가 다시 급격히 감소하고, 350nm 이후로는 차분값이 서서히 감소하는 경향을 보였다.

3.3. 최적변수구간 선정

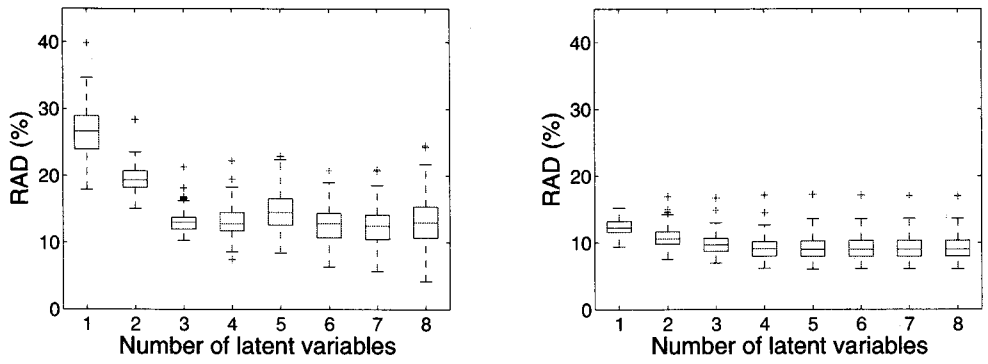
스펙트럼데이터는 어느 영역의 파장을 이용하든 충분한 영역에서 실험하는데 실제 품질변수의 예측모델 개발은 전체영역보다 일부 구간의 흡광도만을 사용할 때 예측정확도가 높아질 수 있다. 선택되는 구간의 특성을 보면 예측하고자 하는 품질변수의 물리적 정보를 갖고 있는 피이크 구간인 경우가 많고, 제외되는 영역을 보면 노이즈부분으로 예측변수를 설명해주기보다는 오차를 주는 영역이다. 현장에서 실험구간을 줄이는 것은 실시간 품질예측시스템 구현 시 시간과 경비를 절약해줄 수 있고, 일부구간의 스펙트럼만 사용함으로써 보다 정확한 실험이 이루어질 수 있는 이점도 있다.

본 연구에서는 VIP값이 특정값 이상인 변수만을 유의한 독립변수로 선택하는데, 우선적으로 예측모델의 최적잠재변수 개수를 결정하기 위해 five-fold cross validation을 수행하고, 선정된 잠재변수를 이용하여 VIP를 계산한 다음 상대절대값오차를 척도로 유의한 변수구간의 선택유무를 결정한다. 예측정확도의 오차척도는 MAD(Mean Absolute Deviation), MSE(Mean Squared Error) 등 여러 가지를 사용할 수 있는데, 본 연구에서는 RAD(Relative Absolute Deviation)를 사용하였다. 품질변수의 경우 현장에서 공정관리를 할 때 공정오차의 허용범위를 설정하고 실제값에 비해 몇%의 오차가 발생하고 있는지를 비교하는 것이 중요하기 때문이다.

$$\mathbf{RAD} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right|. \tag{3.1}$$

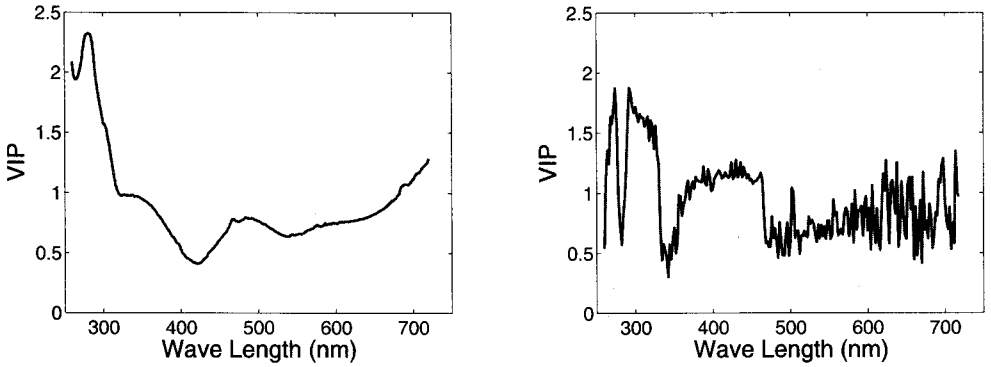
부분최소자승법의 예측모델에서 최적잠재변수개수를 결정하기 위해서 다음과 같은 데이터를 랜덤하게 구성하였다. 코팅두께를 예측하기 위한 시료가 제한적인 이유로 14개 샘플에 대해 4번 반복 실험하였는데, 매 실험마다 흡광도의 변동이 있으므로 이를 랜덤화하기 위하여 각각의 set은 14개의 샘플 시료 각각에서 네 개의 스펙트럼 중 하나씩을 임의로 추출하여 총 100개의 validation set을 구성하고, 각각 five-fold cross validation을 수행하여 100개의 RAD값을 얻는다. 그 다음 100개 RAD값의 평균이 가장 작은 잠재변수 개수를 최적잠재변수개수로 선정하고, 이보다 작은 잠재변수 중 one-sided paired t-test 결과 95% 신뢰수준에서 RAD값이 유의한 차이를 보이지 않는 잠재변수가 있다면, 그 중 최소값을 최적잠재변수개수로 다시 선정한다. 이러한 과정을 통해서 1차 차분 전 데이터에 대해 잠재변수 6개, 1차 차분 후 데이터에 대해 잠재변수 4개를 최적잠재변수개수로 선정하였다 (그림 3.2).

선정된 최적잠재변수를 적용하여 VIP값을 구한 결과, 그림 3.3과 같이 인접한 파장의 VIP값 간의 자기상관이 매우 높은 경향을 보였다(1차 자기상관계수: 1차 차분 전 0.98, 1차 차분 후 0.78). 이처럼 자기



(a) 1차 차분 전 cross validation 오차 (b) 1차 차분 후 cross validation 오차

그림 **3.2.** 잠재변수의 수에 따른 PLS regression 모형의 cross validation 오차 추이



(a) 1차 차분 전, 잠재변수 6개 (b) 1차 차분 후, 잠재변수 4개

그림 **3.3.** 파장별 VIP값

상관이 높은 스펙트럼 데이터에서는, VIP가 높은 일부 구간만을 측정함으로써 측정비용을 줄이는 동시에 예측 정확도를 보장할 수 있다.

코팅두께 예측을 위한 UV 흡광도 스펙트럼의 변수선택은 일반적인 독립변수를 개별선택하는 것이 아니라 일정변수구간을 선택하는 과정이다. 일정파장간격으로 연속적으로 흡광도를 얻기 때문에 스펙트럼 데이터에 있어서 변수선택은 변수구간선택문제이며, 가장 유의한 영향을 갖는 연속된 구간을 선택해야 한다는 제약조건이 있다. 예를 들어, 그림 3.3(a)의 경우, VIP값 1을 기준으로 주요변수를 선정할 때 두 파장구간(260~318nm, 684~720nm)이 선택되는데, 측정 설비는 이 두 구간의 흡광도만을 선택적으로 측정하는 것이 어려우므로 260~720nm 구간을 모두 측정하고 예측모델식에서만 선택된 구간을 사용하게 된다. 현장의 제약조건을 고려한 최적변수구간 선택방법은 다음과 같다.

- Step 1. VIP값이 θ 이상인 파장을 추출한 뒤, 연속된 구간끼리 모아 구간을 구성한다.
- Step 2. 인접한 두 구간 사이의 떨어진 거리가 d_1 이하일 경우, 두 구간을 합하여, 앞 구간의 시작점부터 뒤 구간의 끝점까지가 하나의 구간이 되도록 한다.
- Step 3. 구간의 길이가 d_2보다 짧은 구간은 제거한다.

표 **3.2.** 최적변수구간 선정의 예($\theta = 1$, $d_1 = 6$nm, $d_2 = 14$nm, $v = 0$)

변수구간(nm)	측정구간((총길이), nm)	잠재변수	RAD(평균 ± 표준편차, %)
262 ~ 276	262 ~ 278 (16)	2	11.58 ± 1.93
288 ~ 330	288 ~ 332 (44)	4	6.93 ± 1.57
366 ~ 464	366 ~ 466 (100)	1	21.66 ± 1.61
262 ~ 276, 288 ~ 330	262 ~ 332 (70)	2	6.72 ± 1.02
262 ~ 276, 366 ~ 464	262 ~ 466 (204)	2	7.60 ± 1.38
288 ~ 330, 366 ~ 464	288 ~ 466 (178)	3	7.58 ± 1.53
262 ~ 276, 288 ~ 330, 366 ~ 464	262 ~ 466 (204)	3	6.25 ± 1.58

- Step 4. 구간 내 변수들의 VIP값의 평균에서 θ를 뺀 값이 v보다 작은 구간은 제거한다.
- Step 5. 구간의 조합을 통해 가능한 모든 변수구간을 생성한다.
- Step 6. 각 변수구간을 이용하여 예측모델을 생성하고, cross validation을 통해 가장 예측오차가 작은 구간을 최적구간으로 선정한다. 변수구간 간에 유의한 예측오차 차이가 발견되지 않을 경우, 측정구간이 짧은 것을 우선시한다.

그림 3.3(b)의 경우, $\theta = 1$, $d_1 = 6$nm, $d_2 = 14$nm, $v = 0$일 때 Step 4 종료 후 세 개의 변수구간(262~276nm, 288~330nm, 366~464nm)이 남으며, Step 5에서는 앞의 세 구간 외에 추가로 네 개의 변수구간을 생성하여, Step 6에서 총 7개의 변수구간에 대한 평가를 하게 된다 (표 3.2). 이 예에서 보면, 260~464nm의 총 204nm 구간에서 스펙트럼을 측정하여 262~276nm, 288~330nm 및 366~464nm 구간 데이터를 이용하는 경우에 RAD값이 가장 낮았으며, one-sided paired t-test를 통해 이보다 짧은 측정구간을 지니는 예측모형 중 95% 신뢰수준 하에서 RAD 측면에서 통계적 차이를 보이지 않는 모형이 없으므로, 표 3.2의 7개 변수구간 중 가장 마지막 구간이 최적구간이 된다.

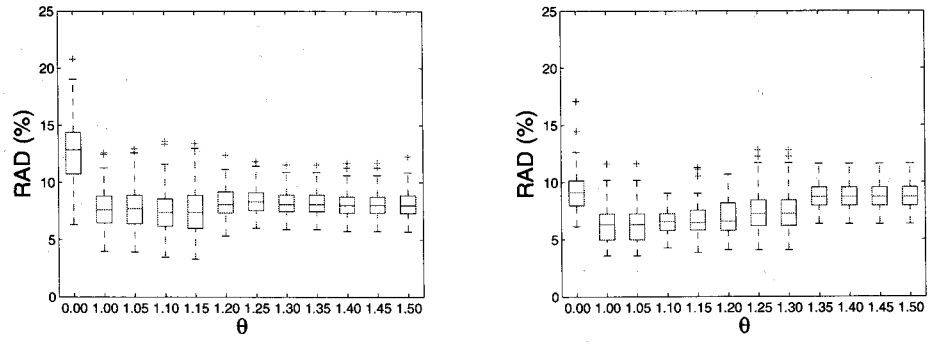
변수구간 선정 방법을 단계별로 상세하게 설명하면, Step 3에서는 d_2값 설정 시, 구간 내 파장의 수가 PLS 최대잠재변수의 수보다 적으면 회귀모형 수립이 불가능하므로, $d_2 \geq 2 \times$ (최대잠재변수 개수 $-$ 1)nm가 되도록 설정한다. 모델생성 시 최대잠재변수 개수를 8개로 잡았으므로 $d_2 = 14$nm로 설정하였다. Step 4에서는 d_1값이 너무 작으면 평가할 구간이 많아져 계산량이 증가하고, 반대로 너무 크면 중요하지 않은 변수가 예측모형에 포함되어 예측력을 떨어뜨리게 되므로, VIP값의 경향에 따라 적절하게 설정하도록 한다. 따라서 구간 내 변수들의 VIP평균값이 θ이하인 조건, 즉 v값이 0보다 작으면 해당 구간은 제거한다. θ값이 작을수록 Step 1에서 더 많은 변수가 선택되며, $\theta = 0$인 경우에는 전체 변수가 선택된다.

본 연구에서는 θ를 1.0부터 1.5까지 0.05 단위로 증가시키면서, 각각의 θ값에 대해 최적변수구간을 선정하고, $d_1 = 6$nm, $d_2 = 14$nm, $v = 0$으로 설정하였다. 이를 변수선택 이전의 예측모델($\theta = 0$) 성능과 비교하였을 때, 전반적으로 cross validation 오차가 크게 감소하였음을 그림 3.4에서 확인할 수 있다.

각 θ값별로 선정된 최적파장구간 및 cross validation 결과, 1차 차분 전에는 260~312, 704~720nm 변수구간을 이용한 PLS 회귀모형이 가장 적합하게 나타났으며, 1차 차분 이후에는 262~276, 288~330, 366~464nm 변수구간을 적용하는 것이 RAD 관점에서 최적이었다 (표 3.3). 두 모형의 RAD를 비교한 결과, 1차 차분 후의 모형이 스펙트럼 측정 구간이 짧고 예측 정확도 또한 높은 것으로 나타났다.

4. 시뮬레이션

앞의 3.3장에서 선택된 최적변수구간이 얼마나 적합한지를 임의로 추출한 100개의 샘플군에서 그 성능



(a) 1차 차분 전 cross validation 오차 (b) 1차 차분 후 cross validation 오차

그림 **3.4.** θ값별 최적모형의 cross validation 오차

표 **3.3.** θ값별 최적변수구간

1차 차분 전				
θ	변수구간(nm)	측정구간((총길이), nm)	잠재변수	RAD(평균 ± 표준편차, %)
0.00	260 ~ 720	260 ~ 720 (460)	6	12.66 ± 2.59
1.00	260 ~ 318, 684 ~ 720	260 ~ 720 (460)	5	7.72 ± 1.70
1.05	260 ~ 316, 686 ~ 720	260 ~ 720 (460)	5	7.79 ± 1.75
1.10	260 ~ 314, 700 ~ 720	260 ~ 720 (460)	6	7.56 ± 2.04
1.15	260 ~ 312, 704 ~ 720	260 ~ 720 (460)	6	7.45 ± 2.10
1.20	260 ~ 312	260 ~ 312 (52)	3	8.26 ± 1.31
1.25	260 ~ 310	260 ~ 310 (50)	2	8.40 ± 1.14
1.30	260 ~ 308	260 ~ 308 (48)	2	8.21 ± 1.12
1.35	260 ~ 308	260 ~ 308 (48)	2	8.21 ± 1.12
1.40	260 ~ 306	260 ~ 306 (46)	2	8.12 ± 1.12
1.45	260 ~ 306	260 ~ 306 (46)	2	8.12 ± 1.12
1.50	260 ~ 304	260 ~ 304 (42)	2	8.22 ± 1.14
1차 차분 후				
θ	변수구간(nm)	측정구간((총길이), nm)	잠재변수	RAD(평균 ± 표준편차, %)
0.00	260 ~ 718	260 ~ 720 (460)	4	9.17 ± 1.79
1.00	262 ~ 276, 288 ~ 330, 366 ~ 464	262 ~ 466 (204)	3	6.25 ± 1.58
1.05	262 ~ 276, 288 ~ 330, 366 ~ 464	262 ~ 466 (204)	3	6.25 ± 1.58
1.10	262 ~ 276, 288 ~ 330, 388 ~ 450	262 ~ 452 (190)	2	6.54 ± 1.09
1.15	288 ~ 330, 424 ~ 446	288 ~ 446 (158)	4	6.68 ± 1.39
1.20	288 ~ 330	288 ~ 332 (44)	4	6.93 ± 1.57
1.25	290 ~ 328	290 ~ 330 (40)	4	7.44 ± 1.75
1.30	290 ~ 328	290 ~ 330 (40)	4	7.44 ± 1.75
1.35	290 ~ 326	290 ~ 328 (38)	2	8.79 ± 1.13
1.40	290 ~ 326	290 ~ 328 (38)	2	8.79 ± 1.13
1.45	290 ~ 326	290 ~ 328 (38)	2	8.79 ± 1.13
1.50	290 ~ 326	290 ~ 328 (38)	2	8.79 ± 1.13

을 비교하였다. 본 실험에서는 제안한 변수구간선정방법이 노이즈가 어느 정도 있을 때 얼마나 강건한 예측결과를 주는지 비교하기 위해 앞 장에서 사용된 56개 데이터에 임의로 노이즈를 첨가하여 새로운

표 4.1. 시뮬레이션 결과

전처리	변수구간(nm)	잠재변수	예측오차(RAD, %)				
			$a=0.01$	$a=0.05$	$a=0.1$	$a=0.5$	$a=1$
-	260 ~ 720	6	4.24	5.45	8.47	19.52	19.90
-	260 ~ 312, 704 ~ 720	6	3.97	5.09	7.70	19.48	19.81
1차 차분	260 ~ 718	4	3.28	5.00	8.19	18.78	20.22
1차 차분	262 ~ 276, 288 ~ 330, 366 ~ 464	3	4.06	4.88	7.81	19.16	19.52

스펙트럼 데이터를 아래와 같이 생성하였다.

$$\mathbf{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_{56}]^T,$$

where

$$\mathbf{s}_i = \left(s_i^{(1)}, \ldots, s_i^{(231)}\right)^T, \quad i = 1, \ldots, 56,$$
$$s_i^{(j)} = x_i^{(j)} + \varepsilon_{ij}, \quad i = 1, \ldots, 56,\ j = 1, \ldots, 231,$$

여기에서 ε_{ij}는 원 데이터에 첨가되는 노이즈를 나타낸다. Σ를 스펙트럼 데이터 $\mathbf{X}$의 공분산 행렬(231 × 231)이라 할 때, 본 실험에서는 ε_{ij}는 평균이 영벡터이고 공분산 행렬이 $a\Sigma$인 다변량 정규분포로부터 생성하였다. 이때, 상수 a가 클수록 노이즈가 많이 포함되며, Σ는 동일 샘플 내 스펙트럼데이터의 측정 편차뿐만 아니라 14개 샘플간의 편차를 포함하기 때문에, 동일 샘플 내의 편차만을 나타내는 $a\Sigma$는 Σ보다 작다고 볼 수 있다. 따라서 본 실험에서는 a의 수준을 0.01, 0.05, 0.1, 0.5, 1 등 다섯 개 수준으로 나누어 노이즈를 생성하였다.

노이즈가 첨가되지 않은 기존 56개 스펙트럼을 이용하여 예측모델을 생성하고, 이 모델에 노이즈가 첨가된 데이터에 적용하여 코팅두께를 예측한 결과가 표 4.1이다. 첨가된 노이즈의 크기가 클수록 예측오차가 증가하였으며, 1차 차분 전후의 예측오차 차이는 미미하였으나, 1차 차분 처리를 한 경우 56% 가량 축소된 스펙트럼 측정구간 하에서 1차 차분 이전의 최적 모형과 동일한 수준의 성능을 나타냈다. 표 3.3의 결과보다 오차가 적은 이유는 56개 데이터 전체로 예측모델을 형성하고 노이즈 첨가한 56개 데이터를 테스트데이터로 사용했기 때문이다. 1차 차분 처리를 한 경우 중 전체변수를 사용한 경우와 일부 변수구간만을 선택적으로 사용한 경우를 비교하였을 때, 노이즈가 매우 적은($a = 0.01$) 경우를 제외하면 두 모델이 대등한 성능을 보였다. 따라서 1차 차분 처리 및 변수구간 선택을 적용한 회귀모형이 예측정확도 및 측정비용 측면에서 가장 적합한 것으로 보인다.

5. 결론

본 연구에서는 스펙트럼데이터를 이용한 품질변수예측에서 데이터전처리 및 변수구간선택방법의 적용이 예측정확도에 미치는 영향을 비교분석하였다. 독립변수로 사용되는 흡광도는 연속된 파장구간에서 나오는 값이므로 어느 특정파장의 흡광도를 선택하는 문제라기보다는 연속된 구간에서의 변수선택문제이다. 본 연구에서 제안한 변수중요도척도를 이용한 변수구간선정방법은 데이터전처리 전의 경우 전체구간을 사용하는 것에 비해 RAD가 12.66%에서 7.45%로 감소하는 효과가 있었으며, 데이터전처리로서 1차 차분을 적용한 경우는 전체구간사용 시 RAD가 9.17%였는데 변수구간선정방법 적용 시 6.25%까지 오차를 감소시키는 효과가 있었다. 코팅두께의 예측모델개발에서는 연속된 구간선정이라는 제약조건이 있어서 인접한 구간이 가까운 경우는 합치고 일정구간보다 좁은 경우는 제거하는 방식으로 최적구간을 선정하는 방법을 적용하여 예측정확도를 향상시킬 수 있었다. 일반적인 공정데이터의 변수

선택에서도 변수중요도척도를 적용한 변수선택은 예측정확도를 향상시킬 수 있을 것이며, 고차원데이터의 경우에는 변수중요도척도로 예측변수에 중요한 영향을 갖는 공정요인을 도출하는 방법으로도 적용될 수 있을 것이다.

참고문헌

Chong, I. and Jun, C. (2005). Performance of some variable selection methods when multicollinearity is present, *Chemometrics and Intelligent Laboratory Systems*, **78**, 103–112.

Cramer, J. A., Kramer, K. E., Johnson, K. J., Morris, R. E. and Rose-Pehrsson, S. L. (2008). Automated wavelength selection for spectroscopic fuel models by symmetrically contracting repeated unmoving window partial least squares, *Chemometrics and Intelligent Laboratory Systems*, **92**, 13–21.

Heise, H. M., Damm, U., Lampen, P., Davies, A. N. and McIntyre, P. S. (2005). Spectral variable selection for partial least squares calibration applied to authentication and quantification of extra virgin olive oils using fourier transform raman spectroscopy, *Applied Spectroscopy*, **59**,1286–1294.

Lee, D., Lee, H., Jun, C-H. and Chang, C. H. (2007). A variable selection procedure for X-ray diffraction phase analysis, *Applied Spectroscopy*, **61**, 1398–1403.

Wold, S., Ruhe, A., Wold, H. and Dunn III, W. J. (1984). The collinearity problem in linear regression. The partial least squares(PLS) approach to generalized inverses, *SIAM Journal on Scientific and Statistical Computing*, **5**, 735–743.

Wold, S., Sjöström, M. and Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics, *Chemometrics and Intelligent Laboratory System*, **58**, 109–130.

A Prediction Model for Coating Thickness Based on PLS Model and Variable Selection

Hyeseon Lee[1] · Youngrok Lee[2] · Chi-Hyuck Jun[3] · Jae-Hwa Hong[4]

[1]Department of Industrial and Management Engineering, POSTECH
[2]Department of Industrial and Management Engineering, POSTECH
[3]Department of Industrial and Management Engineering, POSTECH
[4]Instrumentation Research Group, Technical Research Laboratory, POSCO

(Received January 2010; accepted February 2010)

Abstract

Coating thickness is one of target variables in quality control process in steel industry. To predict coating thickness and to control quality of anti-fingerprint steel coils, ultraviolet-visible spectra are measured. We propose a variable-interval selection procedure based on the variable importance in projection in partial least square model. Using the proposed variable interval selection method, prediction performance gets better in the reduced model than the full model with full spectra absorbance. It is also shown that the first differencing as a data preprocessing technique does work well for the prediction of coating thickness.

Keywords: Partial least square, variable importance in projection, data preprocessing.

[1]Corresponding author: Research Professor, Department of Industrial and Management Engineering, POSTECH, Pohang 790-784, Korea. E-mail: hyelee@postech.ac.kr