선별된 내적변환을 포함한 PLS를 이용한 소프트센서 설계 및 적용

홍선주, 한종훈, 장근수

포항공대 화학공학과, 공정산업의 지능자동화 연구센터

soft sensor design and application using PLS with selected inner relation

Sun Ju Hong, Chong Hun Han, Kun Soo Jang

Department of Chemical Engineering, Automation Research Center,

Pohang University of Science and Technology

1. 서론

최근에는 지구환경을 보호하기 위해 ISO14000 에 대응하는 것은 기업의 사회적 책임을 수행하는 의미에서도 중요한 목표로 되고 있다. 따라서 각 공장의 플랜트에서도 환경에 미치는 영향을 관리하고 또한 건전한 환경을 계속적으로 유지하게 위해 감시용 분석계가 많이 사용되고 있다. 예를 들어, 화학플랜트에서 부산물로 방출될 수 있는 질소산화물(Nox), 황산화물(Sox) 등의 독소가스의 조기감지 및 정확한 양의 감지는 환경측면뿐만 아니라, 안정된 조업을 수행하는 데 있어서도 매우 중요하며, 배수처리 플랜트에서 바다나 하수 등으로 방출하는 처리수의 배수기준의 엄격한 관리는 깨끗하고 안전한 수질공급 측면에서 중요시 되고 있다. 하지만, 공정에서 부산물로서 배출되는 인체에 유해한 가스나, 회분식 폐수처리시스템의 폐수의 법적 기준 값의 예측은 GC(Gas Chromatograph) 등의 실시간 분석기의 낮은 신뢰성 및 비싼 가격 등으로 인하여 직접적인 실시간 추정이 어려운 실정이다.

따라서, 독소가스, 폐수와 폐기물 등의 좀더 신속하고 신뢰성 있는 정성적, 정량적 예측을 위하여 최근에 많은 환경플랜트에 설치되고 있는 DCS(Distributed Control System)와 PIS(Plant Information System)를 이용하여 공정에 대한 막대한 양의 데이터를 실시간으로 받아들이고 또한 방대한 과거의 데이터(historic data)를 저장함으로써, 플랜트의 과거와 현재에 걸친 많은 양의 상관된 데이터를 이용하여 추정하고자 하는 변수들, 즉 가스, 폐수 또는 폐기물의 정성, 정량 값을 간접적으로 추정하는 방법들이 모색되고 있다.

이처럼, 공정데이터를 이용하여 실시간으로 신속하고 정확하게 추정하기 어렵거나 불가능한 품질변수(가스, 폐수 또는 폐기물 등의 정성, 정량 값)를 예측하는 모델을 소프트센서라고 하며, 공정 데이터를 이용하여 품질변수를 추정하는 모니터링를 통해 이상을 조기에 발견(fault detection), 진단(diagnosis)하는데 응용할 수 있다. 한편, 소프트센서를 이용한 추론제어(inferential

control)는 이미 증류공정이나 생물환경공정 및 고분자공정에 널리 응용되고 있다.[4]

소프트센서 설계는 대상공정에 적합한 추정모델을 선정하는 것으로써 크게 세가지 모델링 접근 방법이 있다. 첫째로는 가장 직접적인 방법으로써 수식적인 모델이 있는데, 이는 추정하고자 하는 변수와 공정변수와의 관계가 명확하게 수식적으로 표현 가능할 때 사용할 수 있는 방법으로써 공정의 지배 방정식이 복잡한 화학공정의 경우에는 제한적으로 사용되고 있다. 한편 전문가 시스템과 인공 신경회로망을 사용하는 등의 지식-기반 접근 방법이 있는데 이들은 공정에 대한 자세한 수식적인 모델을 필요로 하지 않는다는 장점이 있다. 그러나 전문가 시스템에서는 공정에 대해 잘 아는 전문가나 조업자의 경험적 지식이 주관적일 가능성이 많고 전문가들 사이의 커뮤니케이션이 잘 이루어지지 않으면 개발 시간 및 비용이 많이 소요되기 쉽다는 단점이 있다. 또한 인공 신경 회로망을 사용하는 방법은 모델을 구성하기 위해 오류 및 이상들을 포함한 대량의 학습 데이터가 있어야 하나 실제 공정으로부터 이런 대량의 오류, 이상의 데이터를 얻기란 굉장히 어렵다는 단점이 있다.

마지막으로 추정 모델링 접근 방법으로써 다변량 통계적 분석 방법이 있는데, Principal Component Regression(PCR), Partial Least Square or Projection to Latent Structure(PLS)등이 대표적이며 이들은 상관관계가 많은 공정의 데이터를 효과적으로 처리를 하여 공정변수와 추정변수와의 관계를 맺고, 이를 모니터링, 진단 및 제어등에 적용하는 방법이다. 이 방법은 방대한 양의 데이터를 간단한 통계적인 방법을 이용하여 쉽게 모델링할 수 있고, 동시에 공정의 데이터 분석에 이용할 수 있는 도구를 얻을 수 있다는 점에서 매우 유용하다.

다변량 통계적 방법에는 PLS외에도 입력(X)과 출력(Y)의 관계를 맺어주는 Multiple Linear Regression(MLR)과 PCR이 있는데, MLR은 데이터간의 강한 상호 작용에 의해 올바른 해를 얻기 힘들며; PCR은 입력변수의 score 벡터들(또는 PC)이 출력변수와는 독립적으로 입력변수의 연관성만을 고려하여 결정되어지므로 고유값(eigenvalue)이 작아서 회귀모델(regression model)에 선택되어지지 않은 입력변수의 score 벡터가 실제로 출력변수와의 관계에 있어서 중요한 정보를 제공할 수 있다는 단점이 있다.[3] PLS는 공정 변수들간의 redundancy를 해결하고 측정 잡음을 (measurement noise) 없애 주면서 이러한 단점들을 극복하고 있기 때문에, 보다 나은 입력과 출력의 관계를 얻을 수 있어서 근래 들어 많은 이론적 연구와 함께 실제 공정에 적용이 중요한 현안이 되고 있다.

현재 가장 많이 쓰이고 있는 PLS 알고리즘은 크게 외적변환(outer relation)과 내적변환(inner relation)으로 이루어져 있는데, 데이터공간에 대해 분포가 넓은 축부터 시작하여 순차적으로 서로 직교하는 a개의 PC를 구하는 Nonlinear Iterative Partial Least Squares(NIPALS) 알고리즘을 이용하여 외적변환(outer relation)을 하며, 외적변환에서 계산되어진 입출력 score 벡터는 선형회귀(linear regression)를 통하여 내적변환을 갖게 된다.[1]

이러한 선형 PLS 알고리즘은 강한 비선형성, 복잡성과 불확실성을 가지고 있는 공정의 추정

모델에는 적합하지 않게 됨으로써, 외적변환을 거쳐 나온 입출력 score 벡터를 대표적인 비선형 모델링 방법인 신경회로망 등의 비선형 회귀를 통하여 회귀의 성능을 높여 모델의 performance를 높이고자 하는 시도가 1990년대부터 활발하게 연구 되고 있다.[2] 하지만, 신경회로망 PLS는 비선형이 강하지 않은 i번째 단계의 PC의 입출력 score 벡터의 관계를 과도하게 학습시킬 우려가 있으며, 이것은 새로운 데이터가 들어왔을 때 overfitting를 야기시키는 요인이 된다.

이 논문에서는 외적변환을 통하여 계산된 입출력 score 벡터의 회귀모델을 통계적 수치인 PRESS를 기준으로 선형/비선형 회귀를 선별하여 새롭게 PLS를 구성하고, 이 알고리즘을 이용하여 C6/C7 splitter column의 탑저 생산품인 톨루엔의 농도를 정량적으로 추정하는 소프트센서 설계에 적용하였다.

2. 이론

PLS

PLS는 PCR방법이 변형 개선된 것으로서 PCR이 입력(X) 데이터간의 direction만 고려한 것에 반해 PLS는 출력(Y) 데이터 행렬의 direction이 품질변수와 가장 큰 covariance를 갖도록 재배열하는 방법이다. 그림 1은 PLS의 원리를 나타내고 있으며 식으로 나타낼 경우에는,

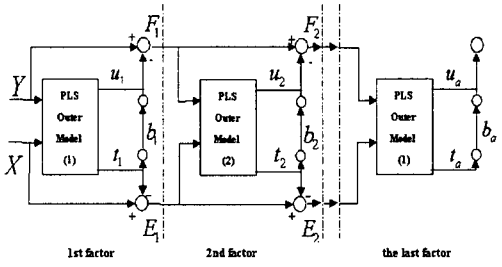

그림 1. NIPALS 알고리즘을 이용한 PLS의 원리

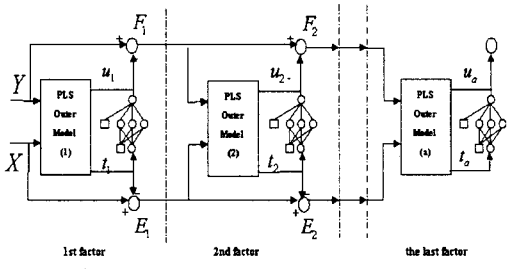

그림 2. NIPALS 알고리즘을 이용한 비선형 PLS의 원리

$$X = TP^T + E \quad (1)$$

$$Y = UC^T + F \quad (2)$$

$$U = BT + R(B = (T^TT)^{-1}T^TU) \quad (3)$$

이다. X는 공정변수인 측정변수를 가리키며, Y는 추정하고자 하는 추정변수에 속한다. PLS 모델은 X 블록의 scores와 Y 블록의 scores가 가장 간단한 형태의 regression으로 이루어져 있으며, 각각 블록 안에서의 외적변환(outer relation)과 두 블록의 관계를 위한 내적변환의 두개의 구조를 이룬다. 식 1), 2)는 각각 X, Y블록에 대한 외적변환을 나타내며 각각의 score t와 u간의 관계가

바로 내적변환을 의미한다. 그리고 여기서 E, F, R은 잔차 행렬(residual matrix)들이고 E와 F가 최소가 되어 거의 공정에 대한 information이 없다고 판단되어 질 때까지(위 그림에서 last factor인 a가 계산되어 질 때까지), 위와 같은 과정이 반복 계산되어 진다. [1]

신경회로망을 이용한 비선형 PLS

식(1)과 (2)의 PLS 외적관계는 그대로 이용하면서 식(3), 즉 내적변환을 다음과 같이 신경회로망을 이용해 비선형 근사를 하는 것이다.

$$u_h = N(t_h) + r_h$$

$N(.)$은 신경회로망에 의해 표현된 비선형 관계를 나타내며, 비선형의 구조는 위와 같다(그림 2).

이때 사용되는 신경회로망의 종류는 multilayer feedforward network, radial basis functions,또는 recurrent networks이 될 수 있다.

이 방법은 PLS 외적변환에 의해 데이터가 전처리가 된 후에 network에 의한 학습이 이루어지므로 비선형 PLS의 network은 SISO이므로 가중치(weights)수가 현저하게 줄어듦으로써 over-parameterization을 피할 수 있을 것이며 또한 국부 최소값의 수도 network의 크기가 작으므로 더 줄어들 것으로 기대할 수 있다.[2]

PLS with selected inner relation

NIPALS 알고리즘을 통하여 입출력 데이터의 외적변환이 계산되고 나면 입출력 score 벡터의 내적변환을 맺게 되는데, 이때 각 단계마다 선형/비선형 회귀모델링을 적절하게 선택함으로써(그림 3) 좀더 강건하고 정확한 모델을 구축하게 된다.

이때 각각의 내적변환 단계에서 PRESS를 통해 선형 또는 비선형 회귀모델을 선택하게 되며, 모델이 구성된 후 예측은 각각의 PC에 저장되어 있는 계수(coefficient)나 가중치(weight)를

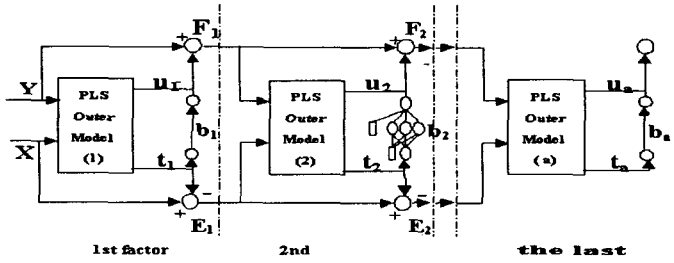

그림 3. PLS with selected innerrelation

이용하여 순차적으로 계산한다.

3. 사례연구

C6/C7 splitter process 개요

Splitter는 60단을 가지고 C6와 C7을 분리해내는 공정이다. 현재 탑저의 톨루엔 조성을 제어하기 위해 42단의 온도를 제어변수로 사용하고 있으며 환류량과 재비기의 열용량이 조작 변수로 사용되고 있다. 측정되는 변수로는 탑상 압력, 탑저 압력, 탑상, 13단, 42단, 56단, 탑저의 온도, 그리고 원료 (Feed), Distillate, Bottom flow rate 등의 13 개이다.

각 데이터는 하루에 두 번씩 8시간과 16시간의 간격을 두고 30분 동안의 측정 치들을 평균한 값이며, 2개월 동안의 조업 데이터로서 Startup을 하면서 Target 조업 조건까지 총 138개의 데이터 중 비 정상 조업 데이터를 제거한 66개의 Observation로 모델을 구성하였다. 그 중에서 56개는 모델을 만드는데 사용하였으며 나머지 10개는 모델의 Prediction Performance을 위해 사용되었다.

모델링 결과 및 고찰

모델이 실제 값의 variance를 설명하는 정도를 R^2를 이용하여 평가하였고, 모델검증을 위하여 시험 데이터와 실제 값의 Mean Square Error of Prediction(MSEP)으로써 각 모델의 prediction power를 비교하였다.(그림 5, 그림 6)

$$R^2 = SSR / SST\,(0 \sim 1),\quad SSR = \sum_{i=1}^{n}(\hat{y} - \bar{y})^2 \quad SST = \sum_{i=1}^{n}(y - \bar{y})^2,\quad MSEP = \frac{1}{n}\sum_{i=1}^{n}(y - \hat{y})^2$$

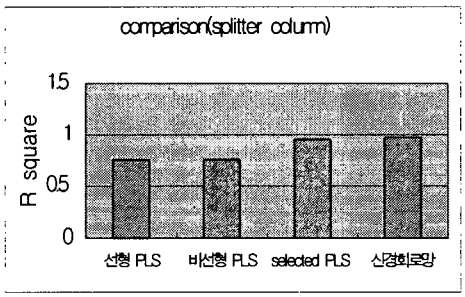

그림 5. 각 모델링에 따른 R^2 비교

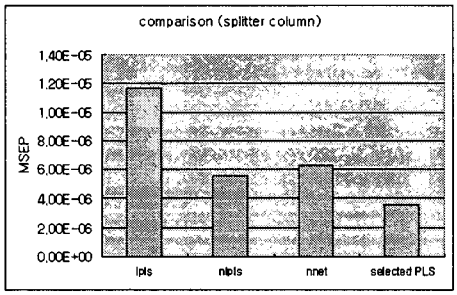

그림 6. 각 모델링에 따른 MSEP 비교

모델이 실제 값의 variance를 설명하는 정도는 신경회로망이 가장 뛰어나지만(그림 5), 모델검증에 있어서 mean square error of prediction이 가장 적은 모델링 방법은 선형 회귀와 비선형 회귀가 적절하게 조합된 PLS with selected inner model이다.(그림 6)

이것은 증류 공정의 비선형 동적 거동성으로 인하여 신경회로망이 모델 데이터를 추정하는 능력은 가장 뛰어나지만 모델 데이터의 수가 충분히 크지 않고 정상조업 데이터의 가능한 variation이 모두 포함되어 있지 않음으로 인해서 시험 데이터의 추정에서 overfitting이 발생한 것으로 보인다. 또한 비선형 PLS도 신경회로망과 비슷한 문제점을 가지게 되면서 각 PC의 내적 변환단계에서 과도한 학습으로 인한 overfitting으로 시험 데이터의 예측능력이 좋지 못하다. 한편,

외적변환을 거친 입출력 score 벡터의 적절한 선형/비선형 회귀모델링을 사용했을 때는 모델링 면에서는 신경회로망과 거의 차이가 없지만, 시험 데이터의 예측에서는 가장 정확하게 톨루엔의 농도를 추정 하고 있음을 알 수 있다.

4. 결론

GC 등과 같은 실시간 분석기의 여러 가지 문제점으로 인하여, 실시간으로 측정하기 어렵거나 불가능한 품질변수의 측정을 공정데이터를 이용하여 간접적으로 추정하는 소프트센서는 증류 공정, 고분자 공정, 생물공정 등의 다양한 산업체 공정에 이용되고 있다. 또한, 최근 공장의 환경 규제가 엄격화 되면서 공정의 부산물로 생산될 수 있는 질소산화물이나 황산화물 등과 같은 독소가스의 조기 감지에도 소프트센서가 이용되고 있다.

이 논문에서는 소프트센서의 모델링 방법으로써 선별된 내적변환을 포함하는 PLS를 제안하였는데, 이는 많은 양의 상관된 데이터를 효과적으로 모델링하는 강건한 선형 PLS와 신경회로망을 이용하여 비선형성을 적절하게 다루는 비선형 PLS를 조합한 알고리즘이다. 이 알고리즘은 사례 연구로써 C6/C7 splitter column의 탑저 생산품인 톨루엔 농도추정을 위한 소프트센서 설계에서 선형PLS, 비선형PLS과 신경회로망보다 좀더 강건하고 정확하게 시험 데이터를 추정하고 있음을 입증해 보였다.

5. 감사의 글

본 연구를 위해 포항공대 공정산업의 지능자동화센터를 통해 재정적 지원을 해 주신 한국과학재단에 감사를 드립니다.

6. 참고문헌

1. J. V. Kresta, T. E. Marlin and J. F. MacGregor, "Development of Inferential Process Models Using PLS", Computers chem. Engng., vol.18, no. 7, pp. 597-611, 1994.

2. S. J. Qin and T. J. McAvoy, "Nonlinear PLS Modeling Using Neural Networks", Computers chem. Engng., vol.16, no.4, pp. 379-391, 1992.

3. Neter, J., Wasserman, W. and Kutner, M.H. : "Applied Linear Statistical models", 3rd ed.

4. Ming T. Tham, Gary A. Montague, A. Julian Morris, and Paul A. Lant, "Soft-sensors for process estimation and inferential control", J.Proc.Cont, Vol 1, pp. 3-14, 1991.