

PLS를 이용한 증류 공정의 국부적인 조성 추정 소프트 센서

홍선주 · 허창구 · 한중훈[†]

포항공과대학교 화학공학과, 공정산업의 지능자동화 연구센터
(1998년 10월 15일 접수, 1999년 3월 4일 채택)

Local Composition Soft Sensor in a Distillation Column using PLS

Sun Ju Hong, Chang Ku Hua and Chong Hun Han[†]

Department of Chemical Engineering, Automation Research Center,
Pohang University of Science and Technology
(Received 15 October 1998; accepted 4 March 1999)

요 약

본 연구는 증류 공정에서 원료 조성의 변화와 같은 외란의 영향, 정상 조업 영역이 넓은 경우, 또는 다양한 품질 규격의 추정(estimation) 등으로 인하여 조업 영역이 여러 개로 구분될 때, 각각의 조업 영역에 맞는 국부적인 조성 추정 소프트웨어의 설계에 관한 것이다. 조업 영역의 분류는 linkage based clustering 방법을 이용하여 구분하였으며, 국부 모델로써 partial least squares(PLS)를 이용하였다. 본 연구에서 제안하는 국부 PLS 소프트센서는 원료 조성의 변화를 갖는 모사된(simulated) 이성분 증류 공정의 에탄올 조성 추정과 정상 조업 영역이 넓은 산업체의 splitter 증류 공정의 톨루엔(toluene) 조성 추정에 적용하였다. 국부 PLS 소프트센서는 global PLS, 신경회로망을 기반으로 하는 비선형 PLS, 신경회로망을 이용한 소프트센서와 비교하였을 때, 두 사례 연구에서 모두 뛰어난 조성 추정 성능을 보여주었다.

Abstract – This paper discusses local composition soft sensors in a distillation column with regards to different operation modes, which are caused by effect of load disturbance such as feed composition change, wide operation regions, estimation of various product specification, etc. Linkage_based clustering method and PLS are used for the classification of various operation regions, modeling of local soft sensor respectively. The proposed method was illustrated using estimation of ethanol in a binary distillation column simulated in simulator, HYSYSTM and toluene in an industrial splitter column within a wide operation window. In simulation, process data for sub-operation regions have been collected to reflect the changes of feed composition that is of frequent occurrence in real distillation column. Local PLS soft sensors were compared to global PLS, nonlinear PLS based on neural network and neural network, and shown better performance than other methods.

Key words : Distillation Column, Different Operation Modes, Local Soft Sensor, Linkage Based Clustering, PLS

1. 서 론

증류 공정의 제품조성 제어는 GC(Gas Chromatography) 등과 같은 실시간 분석기기를 이용하여 이루어진다. 그러나, 실시간 분석기기는 측정지연이 크고 분석하는데 시간이 오래 걸려서 측정 신뢰성이 낮으며, 설치 및 보수 비용도 상당히 비싸기 때문에, 증류 공정에서 제품조성의 측정은 조성제어에 있어서 아주 어려운 문제 중의 하나이다. 이러한 문제를 해결하기 위해서 일찍부터 증류 공정의 조성제어에 추론모델(inferential model)을 이용한 간접적인 제어가 이루어져 왔다.

추론 모델은 달리 소프트센서라고도 하며, 추정하려는 품질변수와 구체적인 함수관계를 갖는 공정의 운전 변수들을 선정하여 대상공정에 적합한 모델을 찾는 일련의 과정을 포함한다[1].

소프트센서는 모델의 접근방식에 따라 이론적인 모델과 경험적인 모델로 나뉘어진다. 이론적인 모델은 공정의 거동을 물질 및 에너지 방정식에 근거한 수학적 모델을 기본으로 하며 단순한 형태의 관계식

이 정확하게 찾아질 경우 정교한 추정이 가능하다. 특히 증류탑 추정기에 대해선 개방형 방정식 기반의 비선형 프로그래밍을 이용해 상당히 정확한 추정을 할 수 있음이 입증되었다[2]. 그러나, 일반적으로 실제 공정에는 공정의 복잡성(complexity)과 공정에 대한 정보(information)의 미비함으로 인하여 입출력 데이터를 그 기반으로 하는 경험적 모델이 실제로 많이 사용되고 있다.

데이터를 기반으로 하는 소프트센서는 수집된 데이터만을 이용하여 모델을 설계하기 때문에 데이터의 양과 질이 모델의 정확성과 직접적으로 연관이 된다. 즉, 경험적 모델에 근거한 소프트센서는 외란(disturbance)의 종류 및 크기, 조업운전범위의 예측이 중요하다[3].

증류 공정에 발생할 수 있는 외란의 종류는 원료조성의 변화, 원료 온도 및 유량의 변화 등이 있는데, 이중에 원료 조성의 변화는 수입 원산지의 변화나 폐유의 재사용 또는 원료 탱크의 차이 등에 기인하며 완전히 새로운 조업영역을 야기시키기에 따라, 미리 설계되어 있는 소프트센서의 추정 능력을 저하시켜 규격(specification)미달의 제품(product)생산을 야기시키는 원인이 된다.

또한 공정의 조업 영역이 넓은 경우 수집된 데이터의 covariance구

[†]E-mail : chan@postech.ac.kr

조가 군집(cluster)을 이루며 완전히 달라짐에 따라 기존의 소프트센서의 추정 성능을 저하시키기도 한다.

따라서 새로운 조업 영역을 야기시키는 예상 외란의 종류 및 크기에 따라서 적절하게 대응할 수 있으며, 넓은 정상 조업 영역에도 강건하고 정확한 추정이 가능한 소프트센서 설계가 요구되어지고 있다.

한편 수집된 공정 변수들은 강한 연관성을 가지고 있는데, Kresta [4]는 PLS가 차원이 크고 상관성이 강한 데이터를 효과적으로 다루면서 좋은 예측능력과 공정의 잡음과 센서의 오차 및 고장에도 강건한 결과를 가져온다고 보고하였다.

이에 본 연구에서는 증류 공정에서 load disturbance인 원료 조성의 변화, 넓은 정상 조업 영역과 다양한 품질규격의 추정 등으로 인하여 조업영역이 여러 개로 나뉘어 질 때, 제안하는 linkage based clustering 방법을 이용하여 조업 데이터를 효과적으로 분류하고 분류된 각각의 조업영역에 PLS를 이용하여 국부 조성 추정 소프트센서를 설계하였다. 2절에서는 데이터 기반 소프트센서 설계 일반에 대해서 기술하였고, 3절은 본 연구에 사용된 변수 선정 방법을, 4절에서는 데이터 전처리 단계로써 데이터 조절과정과 linkage based clustering 방법을 기반으로 하는 데이터 분류(classification)에 대해서, 5절은 PLS알고리즘과 추정능력 평가에 대한 기준을, 그리고 6절에서는 사례연구를 통하여 제안한 알고리즘의 효율성을 검토했다.

2. 데이터 기반 소프트센서 설계 일반

데이터 기반 소프트센서 설계는 크게 다음과 같은 순서를 따른다.

- (1) 변수 선정
- (2) 데이터 전처리
 - 데이터 조절(data scaling)
 - 데이터 분류(data clustering)
- (3) 모델링 및 검증

변수 선정단계는 공정의 정상 상태 데이터를 수집하여, 수집된 데이터 중 추정하려는 제품의 조성과 연관되는 공정 운전 변수를 선정하는 단계이다. 조성추정에 중요한 공정 변수가 선정이 되면, 추정모델의 신뢰성 및 정확성을 위하여 데이터 전처리 단계를 거친다.

일반적으로 데이터 전처리는 데이터 조절(scaling)과 정상조업에 벗어나는 특이한 데이터(outlier) 제거가 순차적으로 이루어지는데, 본 연구의 데이터 전처리 단계에서는 수집된 데이터 조절을 마치면, 데이터 분류(3절 참조)가 이루어지고, 데이터 분류과정에서 특이한 데이터가 제거된다.

변수 선정 및 데이터 전처리 단계를 거친 후 PLS방법을 이용하여 추론 모델(inferential model)을 세우고, 모델의 검증은 cross validation을 통해 최소의 Prediction Error Sum of Squares(PRESS)을 갖는 최적의 principal component(PC) 수를 결정하게 된다.

일반적으로 설계된 모델이 모델 데이터 집합을 설명하는 정도는 R^2 를 통해 알아보고, 모델의 추정성능(prediction power)은 MSE(Prediction)로써 판가름한다.

3. 데이터 선정(Selection of Important Variables)

본 연구에서는 측정 가능한 모든 공정 운전 변수와 추정하고자 하는 조성간에 PLS방법을 이용한 뒤 조성 추정에 별로 기여하지 않는 변수를 제외시키는 방법을 사용한다.

구체적으로 이 방법은 모델이 구성된 후 변수들이 갖는 회귀계수(regression coefficients) 중 0값 근방을 갖는 변수들, 즉 조성 추정에

아주 적은 영향만을 끼치는 변수들을 미리 제거시킴으로써 추론 모델을 좀더 강건하게 구성하였다. 이때 회귀계수는 모델에서 잡은 모든 principal component(PC)에 대한 것이며, 조절된 데이터의 상대적인 비교를 통해서 이루어진다.

한편 PLS의 회귀계수는 품질 변수 추정에 중요한 변수들을 선정하는 방법으로 사용될 뿐만 아니라 모델링이 끝난 후 공정변수와 품질변수간의 상관관계를 분석하는데에도 사용될 수 있다.

4. 데이터 전처리(Data Preprocessing)

4-1. 데이터 조절(Data Scaling)

데이터 조절은 각각 다른 단위를 갖는 여러 종류의 변수값들을 적절한 크기로 조절하는 것이다. 본 연구에서는 auto scaling 방법을 이용하여 변수의 크기를 조절하였다.

자료행렬에서 행(row)은 관측 수(observation number)를 나타내고 열(column)은 변수를 나타내는데, 이때 각 열의 평균값을 빼고 그 값을 그 열에 해당하는 표준 편차로 나누어 자료 행렬의 모든 변수들이 평균값 0과 1의 분산값을 갖도록 조절하였다.

식으로 나타내면 다음과 같다.

$$\text{scaled } x_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j} \quad (1)$$

4-2. 데이터 분류(Data Clustering)

Clustering 방식은 크게 hierarchical clustering과 non-hierarchical clustering으로 나뉘어지는데, 전자는 데이터의 계층적 구조를 파악하면서 진행이 되고, 후자는 데이터를 군집(cluster)으로 불리는 subset으로 구별하는 것에 주안점을 둔다. 여기서는 좀더 정확한 소프트센서 모델링을 위한 데이터 분류 과정으로써 non-hierarchical clustering을 이용하였다.

기존의 non-hierarchical clustering에서 많이 이용되는 방법인 K-means clustering 같은 norm based clustering방법은 데이터가 구형의 covariance 구조를 가지지 않거나 각 군집에 해당하는 데이터의 개수가 서로 크게 차이나는 경우 clustering이 제대로 되지 않는 문제점이 있어 화학공정의 데이터를 처리하기에는 부적합한 점이 있었다. 또 데이터의 밀도 함수를 구성하면서 clustering을 하는 방식인 K-nearest neighborhood clustering이나 kernel clustering 방법 등은 앞의 문제를 해결할 수 있지만 밀도 함수의 매개변수(parameter)를 구하기 위해서 비선형 최적화 문제를 풀어야 하는 어려움이 있다[11].

데이터 분류 방법으로 본 논문에서 제안하는 linkage based clustering방법은 non-hierarchical clustering방식으로써 PCA방법을 통한 score chart에서 데이터 분포를 미리 확인한 후, 각 군집에 해당되는 임의의 데이터를 시작점(seed)으로 하여 서로 가까운 임의의 k개의 점을 선정하고 k개만큼 각 데이터 점들간의 상호간에 linkage를 함으로써 군집을 형성하는 방법이며 그 알고리즘은 다음과 같다.

Step 1. 각 데이터 점에 대해 가장 가까운 K개의 점 NN_{ik} 을 찾고 가까운 순서대로 index를 기록한다.

Step 2. PCA(principal component analysis)방법을 이용하여[9] 데이터 차수가 감소된 score 공간에서 군집의 개수(C)를 정하고 각 군집 내의 임의의 데이터를 선정하여 S_c (seed)라고 하며, 다음과 같이 초기화 한다.

$$CP_{11} = S_1, CP_{21} = S_2, CP_{31} = S_3, \dots$$

Step 3. CP_{ij} 의 NN_{ik} (nearest neighbor) 중에 아직 clustering되지 않은 NP_i (non-classified points)을 찾는다.

Step 4. NP는 index에 포함되는 각 군집의 거리의 역수의 합을 최소로 하는 군집에 지정된다. 이때 index에 포함되어 있지 않은 군집은 연산에 포함시키지 않음으로써 모든 군집에 많이 벗어나는 NP는 어떠한 index에도 포함되지 않는 특이한 데이터로 간주되어 제거된다.

Step 5. 전체 데이터 중에서 CP_c 가 $(100-a)\%$ 만큼 clustering되면 Step 6으로 진행하고 그렇지 않으면 Step 3으로 진행한다.

Step 6. 각 군집의 평균 벡터를 계산하고 그 평균 벡터에 가장 가까운 군집 내의 점을 S_c 로 update한다.

Step 7. Step 2-step 6을 반복하고 모든 군집에 변화가 없으면 중지한다.

여기서 제안한 Linkage based clustering 방법은 K-nearest neighborhood clustering 방법을 그 이론적 기반으로 삼아 다양한 covariance 구조나 비선형 구조를 가지는 데이터 집합도 구별이 가능하며 K-means clustering의 center 개념을 도입하여 수렴할 때까지 계속 update함으로써 linkage를 근간으로 하는 방법에서 나타날 수 있는 clustering이 bias되는 것을 방지한다. 또한 상호간의 linkage index를 통해 데이터 중에서 outlier에 해당하는 점들은 어떠한 군집에도 받아들이지 않음으로써 군집이 오염되는 것을 막을 수 있다[12].

5. 모델링 및 검증(Modeling & Validation)

5-1. PLS(Partial Least Squares)

PLS는 PCR(principal component regression)[6] 방법이 변형 개선된 것으로서 PCR이 측정 데이터간의 연관성만 고려한 것에 반해 PLS는 측정 데이터 행렬의 연관성의 방향(direction)이 품질변수와 가장 큰 covariance를 갖도록 재배열하는 방법이다. Fig. 1은 PLS의 원리를 나타내고 있으며 식으로 나타낼 경우에는,

$$X = TP^T + E \quad (2)$$

$$Y = UC^T + F \quad (3)$$

$$U = BT + R \quad (B = (T^T T)^{-1} T^T U) \quad (4)$$

이다. X는 측정변수를 가리키며, Y는 추정하고자 하는 품질변수에 속한다.

PLS 알고리즘은 크게 외적변환(outer relation)과 내적모델(inner model) 단계를 거치게 되는데, 외적변환은 PCA의 대표적 알고리즘인 Non-linear Iterative Partial Least Squares(NIPALS) 알고리즘을 이용하여 다변수 입력력 데이터가 각각 가장 넓게 분포하고 있는 방향부터 순차적으로 새로운 벡터(principal component)를 정의하고 그 벡터에 정사영된 값, 즉 score 벡터를 만드는 과정이다.

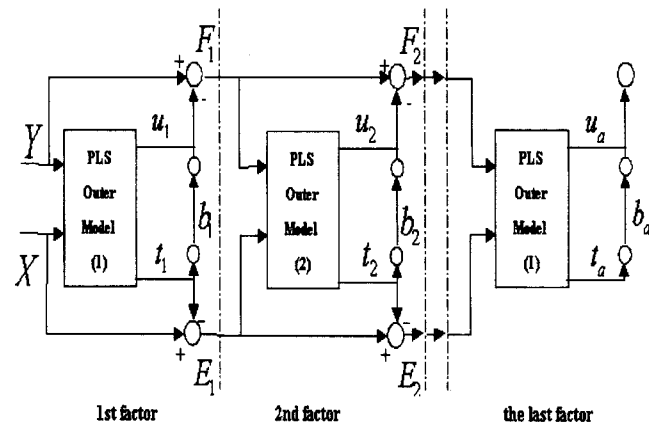


Fig. 1. Principles of PLS.

외적변환을 거쳐 나온 입력력 score 벡터는 최소 자승법(least square)을 이용하여 최종적으로 입력력 데이터의 covariance 구조를 가장 잘 나타낼 수 있도록 재배열되는데 이 과정이 내적 모델링 과정이며, 입력력 데이터의 상관성을 동시에 나타내는 principal component(PC)가 최종적으로 결정된다.

식 (2), (3)은 각각 X, Y블록에 대한 외적변환을 나타내며, 식 (4)는 입력력 score, t와 u간의 관계식으로써 내적 모델(inner relation)을 의미한다. 그리고 여기서 E, F, R은 잔차 행렬(residual matrix)들이고 E와 F가 최소가 되어 거의 공정에 대한 정보가 없다고 판단되어질 때까지(위 그림에서 last factor인 a가 계산되어질 때까지), 위와 같은 과정이 반복 계산되어 진다[6-8].

이때 몇 개의 principal component(PC)를 사용하여 모델링을 할 것인가는 cross validation을 이용하여 결정한다[9].

5-2. Evaluation Criteria

5-2-1. R^2

구성된 모델이 주어진 데이터 집합을 얼마만큼 설명하고 있는가를 나타내는 통계적 지수로써 다음과 같은 식으로 표현된다.

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (5)$$

R^2 값은 0에서 1사이의 값을 나타내며 1로 가까워질수록 모델이 모델에 사용된 데이터를 잘 설명하고 있음을 나타낸다.

5-2-2. MSEP(Mean Squares Error of Prediction)

최종적으로 모델의 성능은 구해진 모델을 새로운 데이터 집합에 이용하여 예측값과 실제값의 차이를 계산함으로써 알 수 있다.

이때 사용한 수치는 MSEP로써 하나의 관측치가 갖는 모델의 오차이다. 식으로 나타내면 다음과 같다.

$$MSEP = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (6)$$

6. 사례 연구(Case Study)

6-1. 이성분 증류 공정의 모사(Simulation of Binary Distillation Column)

본 연구는 원료의 조성변화에 따른 조업 영역이 여러 개로 분류될 때 각각의 조업 영역에 해당하는 국부 소프트웨어 성능을 검증하는 첫 단계로써 이성분 증류 공정에서 탑상의 에탄올 조성을 추정하는

Table 1. Steady-state simulation conditions

	Base condition	Variations in steady state reference set
Inputs		
Feed flow rate	30 m ³ /hr	Constant
Feed temperature	82 °C	80-85 °C
Feed composition		
Ethanol(wt%)	50 %	30-70 %
Water(wt%)	50 %	30-70 %
Outputs		
Distillate		
Ethanol(wt%)	92 %	89-95 %
Bottom		
Ethanol(wt%)	5 %	constant

Table 2. Dynamic simulation conditions

Tray size	
Diameter	11 cm
Weir height	1 cm
Condensor vessel volume	20 L
Reboiler vessel volume	10 L
Tower volume	45 L
Cooling volume	10.2 L
Liquid holdup time	9 min
Setpoint change	
Top	0.89→0.92→0.90
Bottom	Constant

소프트센서 설계에 관한 것이다.

예제로 사용된 이성분 증류 공정은 40개의 이론 단수와 완전 응축기(total condenser), 그리고 재비기로 이루어져 있고, 정상 상태와 동적 상태(dynamic state)의 모사는 모사기, HYSYS™를 통해 수행하였고 각각의 조건은 Table 1, 2와 같다.

Table 1에 주어진 모사조건으로부터 각각 50개의 원료의 에탄올 조성 변화(40% → 50% → 60%)에 따른 기준 데이터를 만들고, linkage based clustering 방법을 이용하여 조성 변화에 따른 각각의 조업 영역에 해당하는 데이터를 분류하였다.

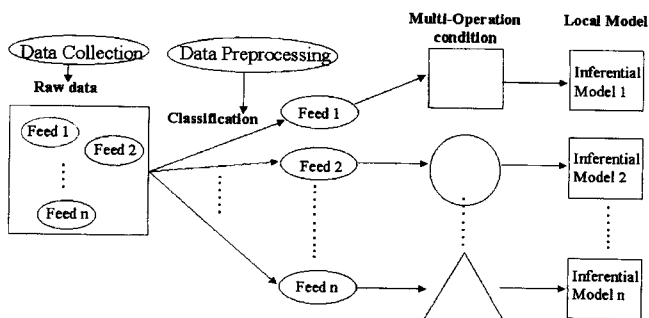


Fig. 2. Proposed procedures of simulation.

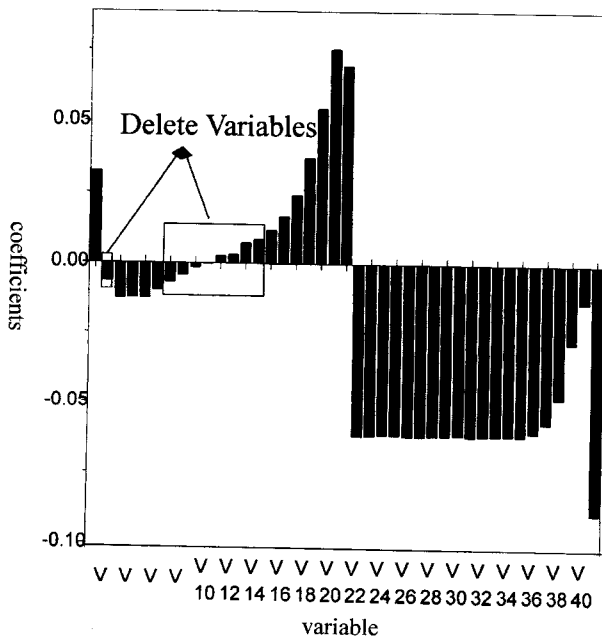


Fig. 3. Selection of important variables.

화학공학 제37권 제3호 1999년 6월

입력 변수는 전단의 온도와 재비기(reboiler)의 온도이며 출력 변수는 탑상의 에탄올 농도이다.

분류된 데이터를 통하여 국부 PLS 소프트웨어를 구성하고, 추정 성능의 비교를 위하여 다른 세 개의 소프트웨어-global PLS, 비선형 PLS, 신경 회로망을 설계하였다. 소프트웨어 설계에 사용되는 데이터는 50개 중에 40개이며, 나머지 10개는 설계된 소프트웨어의 예측성능을 알아보고자 시험 데이터(test data)로써 이용하였다. 이때 비선형 PLS는 McAvoy[10]의 방법을 사용하였으며, 신경 회로망은 모델의 overfitting을 방지하기 위하여 데이터 전처리로서 PCA를 이용한 후 단층 은닉층(hidden layer)으로써 학습하였다. 또한 국부 PLS 소프트웨어의 동적 성능을 비교하기 위해 Table 2로부터 동적 데이터를 얻었다. Fig. 2는 제안하는 알고리즘에 따른 모사 순서에 관한 것이다.

6-1-1. 변수 선정 및 데이터 분류과정

Fig. 3은 변수 선정 결과를 보이고 있다. 탑 하부부터 변수의 번호가 매겨짐에 따라 첫번째 변수는 재비기의 온도를 나타내며 두 번째 변수부터는 증류탑의 단의 온도이다. 이때 bottom의 온도와 밑 단에서 일곱번째에서부터 열네번째 온도까지는 탑상의 에탄올의 조성을 추정하는데 거의 영향을 미치지 않고 있음을 볼 수 있다.

변수 선정이 끝나면, 정적 상태, 동적 상태 각각 150개의 데이터를 수집하여 PCA방법을 통해 차원이 감소된 이차원 공간에서(t score chart) 군집을 이루고 있는 데이터를 확인하고, linkage based clustering방법을 이용하여 데이터를 분류한다. 이때 분류 성능은 SOM, Fuzzy-c mean clustering 방법과 비교하였다.

Fig. 4는 분류 성능에 대한 비교 그림이다. 그림에서 보는 바와 같이 SOM과 Fuzzy-c mean clustering 방법은 그림과 같은 covariance 구조를 갖는 데이터의 경우 잘못 분류하는 데이터의 수가 많다.

Fig. 5의 왼쪽은 정상 상태 데이터가 PCA과정을 거친 후 차수가 감

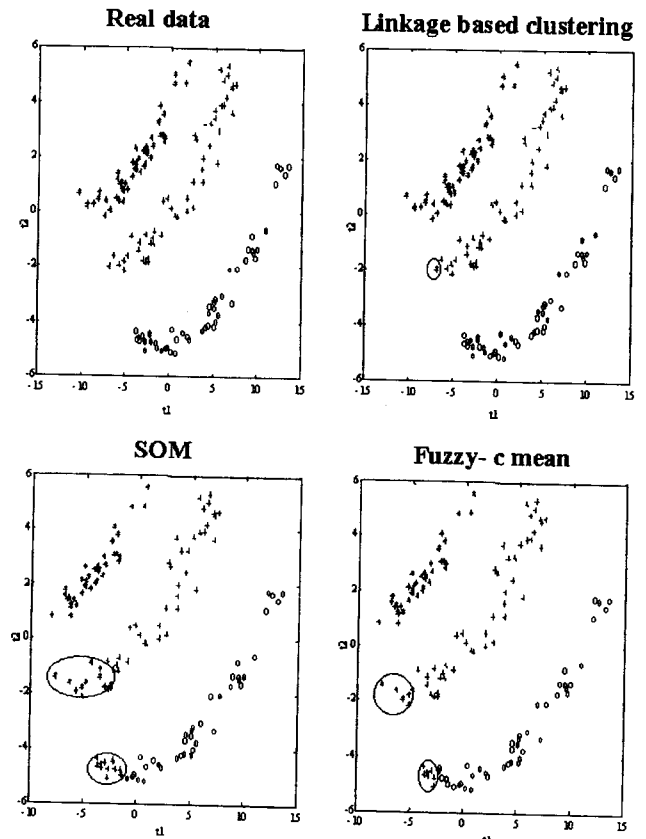


Fig. 4. Comparison of classification power.

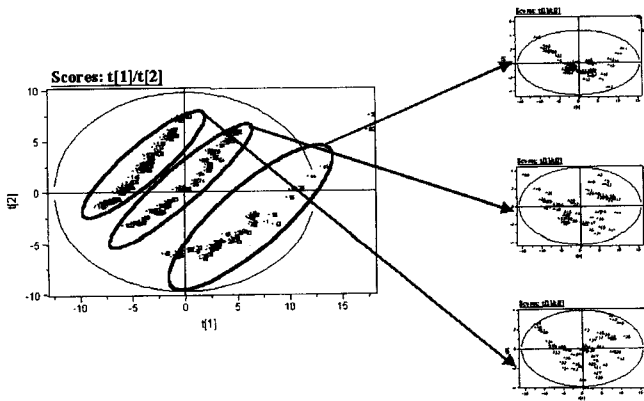


Fig. 5. Data classification.

소된 이차원 공간에서 원료의 조성 변화에 따라 세 개의 군집을 이루고 있는 것을 보여주고 있고, 오른쪽은 데이터 분류를 통한 후 각각의 군집이 갖는 데이터의 이차원 공간을 보여주고 있다.

6-1-2. 정적 성능

모델을 구성한 결과, 국부 PLS는 원료의 에탄올 조성변화(40% → 50% → 60%)에 따라 각각 1개, 2개, 2개, global PLS는 2개, 비선형 PLS는 2개의 PC를 가졌다.

국부 소프트웨어의 정적 성능을 전 조업 영역 구간을 포함하는 여

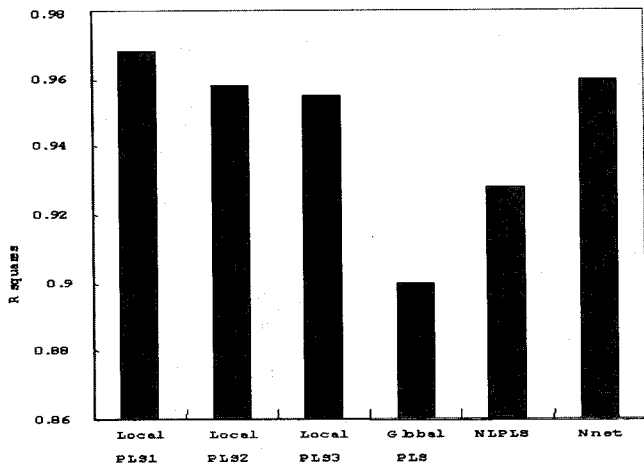
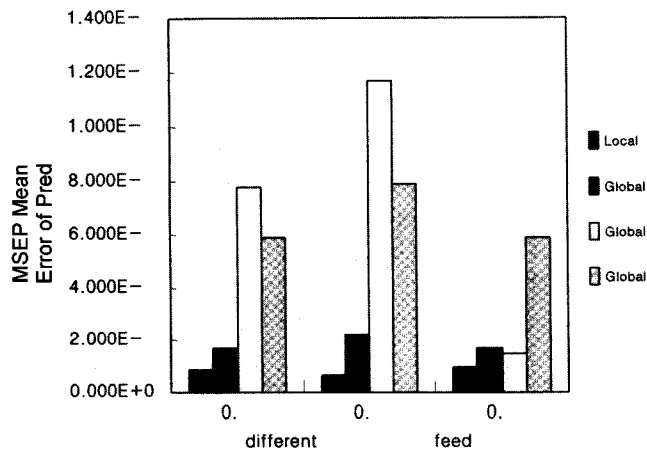
Fig. 6. Comparison of R^2 .

Fig. 7. Comparison of MSEPE between local PLS and global models.

러 소프트웨어의 성능과 함께 비교한 결과가 Fig. 6, 7에 있다.

Fig. 6은 국부 PLS 소프트웨어와 전구간 PLS, 비선형 PLS, 신경 회로망 소프트웨어의 R^2 비교 그래프인데, 모델이 모델 데이터를 설명하는 정도는 국부 PLS 소프트웨어와 신경 회로망 소프트웨어가 가장 뛰어남을 알 수 있다.

Fig. 7은 원료 조성이 변화할 때 국부 PLS 소프트웨어와 다른 종류의 소프트웨어 예측 성능을 시험 데이터의 MSEPE를 이용하여 비교한 그림이다.

결과적으로 국부 PLS 소프트웨어 예측 성능이 가장 뛰어나며, 다른 소프트웨어 중에는 선형 PLS방법을 이용한 소프트웨어가 다른 비선형 모델링 방법보다 예측에 대한 신뢰도가 높았다. 이것은 모사 된 이성분 증류 공정이 운전되는 조업 영역에서 선형성을 띄고 있음을 나타낸다.

6-1-3. 동적 성능

일반적으로 증류 공정의 소프트웨어는 대부분 제어 목적으로 사용되고 있기 때문에 소프트웨어의 동적 성능은 정적 성능과 더불어 중요하다. Fig. 8은 Set point변화에 따른 국부 PLS와 전구간을 포함하는 PLS 소프트웨어의 추정 성능에 대한 비교이다.

Set point의 변화(0.89 → 0.92 → 0.90)에 국부 PLS 소프트웨어는 뛰어난 추정 성능을 보이고 있다.

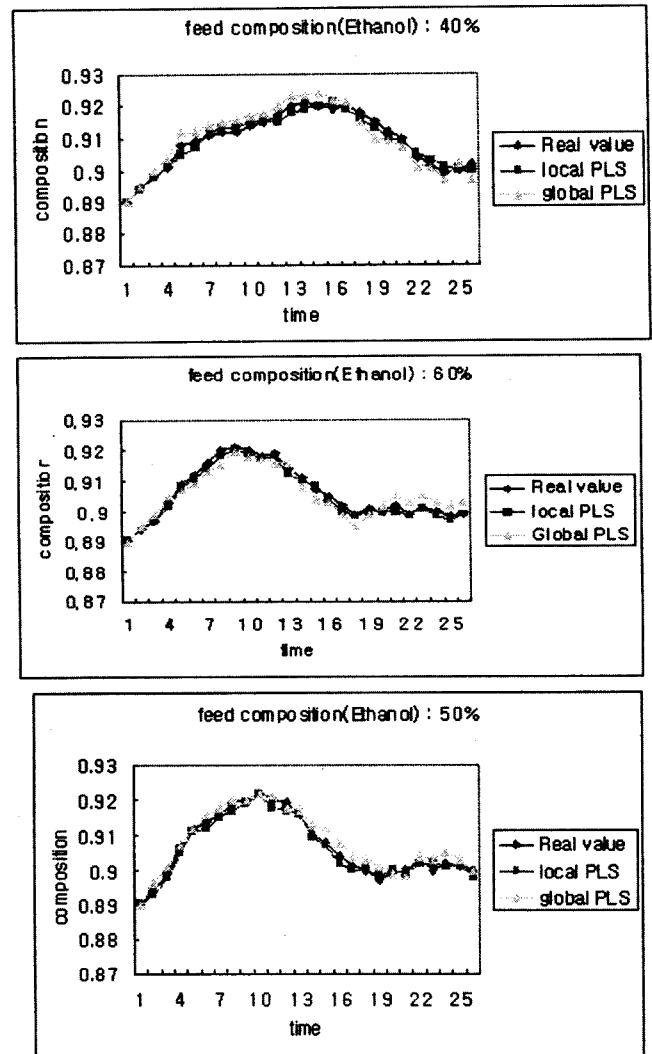


Fig. 8. Comparison of estimation between local PLS and global PLS: set point change(0.89→0.92→0.90), different feed composition.

한편 Set point의 변화가 주어진 조업 영역 안에 존재하고 있기 때문에 전구간의 조업 영역을 포함하는 PLS 소프트웨어의 경우도 대체로 추정 성능이 좋다.

6-2. 산업 splitter 종류 공정(Industrial Splitter Column)

본 사례 연구는 정상상태의 조업 영역이 넓은 탑저(bottom)의 톨루엔(toluene) 조성 규격(specification)을 추정하는 산업 splitter 종류 공정에 국부 소프트웨어를 설계하여 앞선 사례 연구에서 비교 대상으로 사용되었던 여러 종류의 소프트웨어와 추정 성능을 비교하는 것이 목적이다.

본 splitter는 60단을 가지고 C6과 C7을 분리해내는 공정이다. 현재 제품의 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 등이다.

각 데이터들은 하루에 두 번씩 8시간과 16시간의 간격을 두고 30분 동안의 측정치들을 평균한 값이다. 2개월 동안의 조업 데이터로서 정상상태에서의 톨루엔 조성의 범위는 0.02-0.07이며, 개수는 52개다.

수집된 데이터 중에서 42개는 모델을 만드는데 사용하였으며 나머지 10개는 모델의 예측성능을 알아보기 위해 사용되었다.

6-2-1. 변수 선정 및 데이터 분류 과정

Fig. 9는 탑저의 톨루엔의 조성을 추정하는데 기여도가 적은 변수를 PLS모델을 통해 제거하는 과정을 나타내고 있다.

이때 제거되는 변수는 over head의 온도, 탑저의 유량과 재비기의 온도이며, 톨루엔의 조성 추정에 기여를 많이 하고 있는 변수는 탑저의 온도와 원료의 유량, over head의 유량이다.

선정된 변수를 이용하여 수집된 데이터를 PCA방법을 이용하여 차원이 감소된 이차원 공간에서 보았을 때 Fig. 10과 같이 크게 두 개의 군집을 형성하고 있음을 알 수 있다.

오른쪽에 형성된 군집은 조성이 0.02-0.04근방에 있는 데이터이며, 왼쪽에 형성된 군집은 0.04-0.07의 조성을 갖는 톨루엔이다.

한편, 수집된 데이터 중에서 정상 조업 영역을 벗어나는 특이한 데이터가 두 개 있는데, 그림에서 세모로 표시된 오른쪽 위 상단의 데

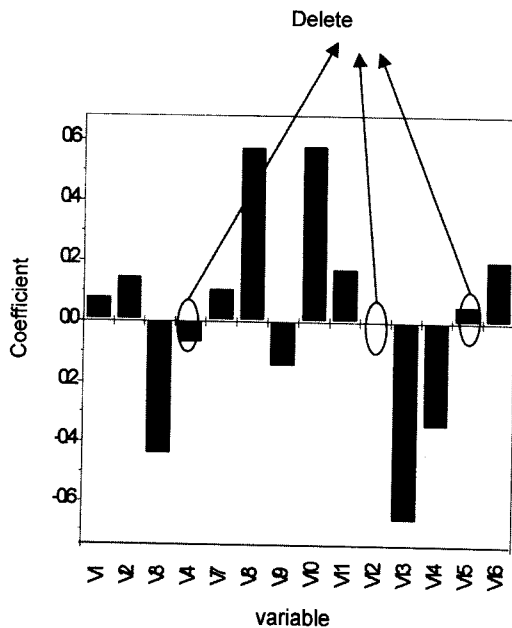


Fig. 9. Selection of important variables.

화학공학 제37권 제3호 1999년 6월

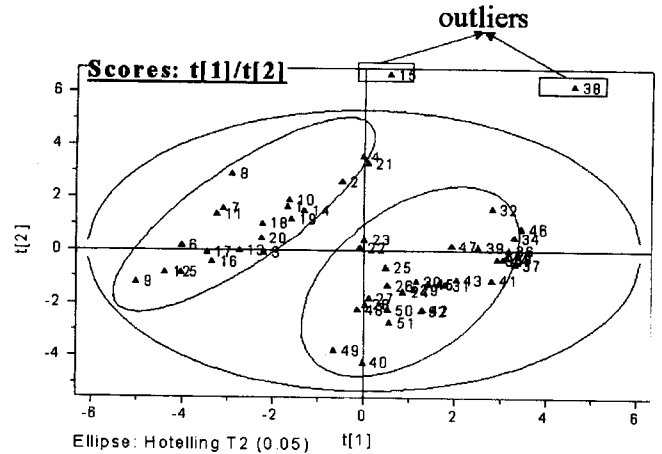


Fig. 10. Preliminary clustering by PCA: two outliers detection.

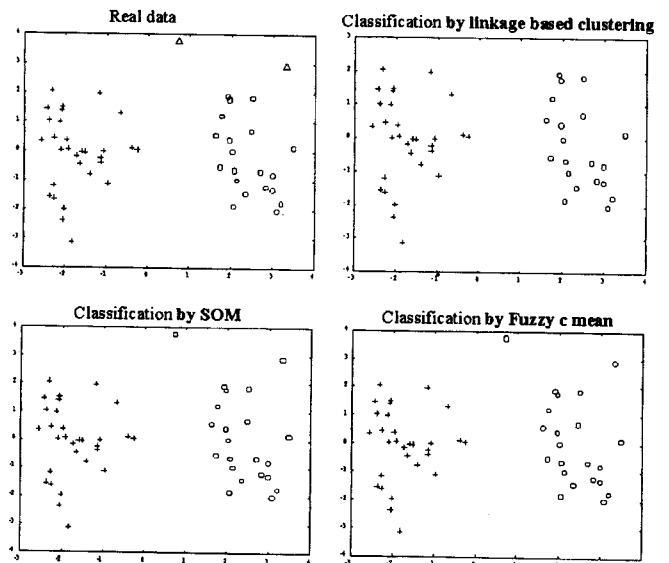


Fig. 11. Comparison of classification power.

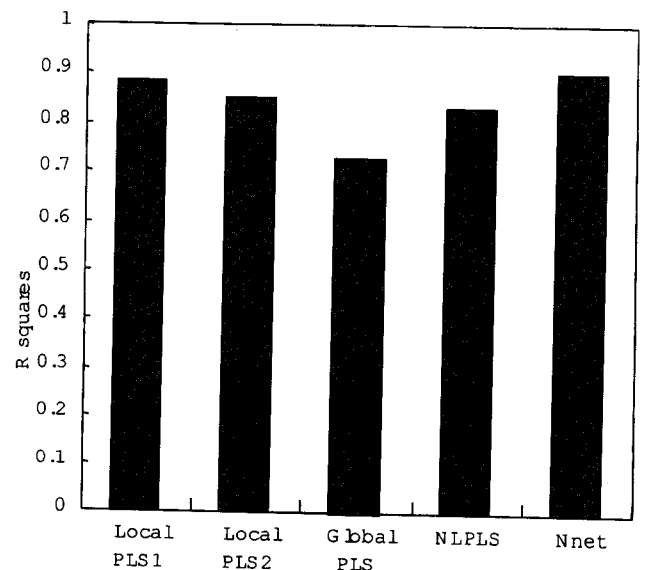


Fig. 12. Comparison of R^2 .

이터가 그것이다.

이 연구에서 데이터 분류 과정은 조업 영역이 다른 두 개의 군집을 분류하고 특이한 데이터를 제거하는 것을 포함한다. Fig. 11은 데이터 분류방법에 대한 비교 그림인데 본 연구에서 제안하는 분류 방법인 linkage based clustering방법(오른쪽 상단)은 데이터 분류를 통하여 조업 영역을 두 개의 군집으로 나누고 있으며 특이한 데이터를 제거하고 있다. 반면, SOM과 Fuzzy-c mean 방법은 둘 다 오른쪽 군집에 특이한 데이터를 포함하고 있다.

6-2-2. 모델링 결과 및 예측성능 비교

변수 선정과 데이터 분류를 마친 후 모델을 구성한 결과 각각 국부 PLS는 5개, PLS는 7개, 비선형 PLS는 4개의 PC를 가졌다.

Fig. 12는 모델의 성능 정도에 대한 R^2 비교 그래프이다. 그림에서 보면 선형 PLS의 R^2 가 가장 낮는데, 이것은 정상 상태의 조업영역이 넓음에 따라 실제 산업 증류 공정의 비선형성이 두드러지고 있음을 나타낸다.

설계된 국부 PLS 소프트웨어의 예측 성능은 선형 PLS 소프트웨어, 비선형 PLS 소프트웨어, 신경회로망 소프트웨어와 함께 Fig. 13, 14에 비교되어 있다.

7. 결 론

본 연구에서는 두 종류의 사례 연구를 통하여 증류 공정에서 원료

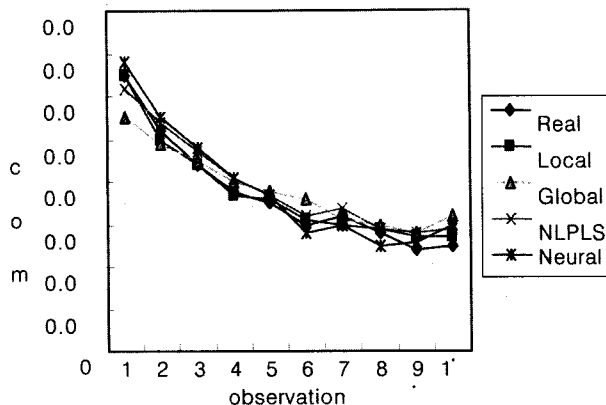


Fig. 13. Comparison of estimation.

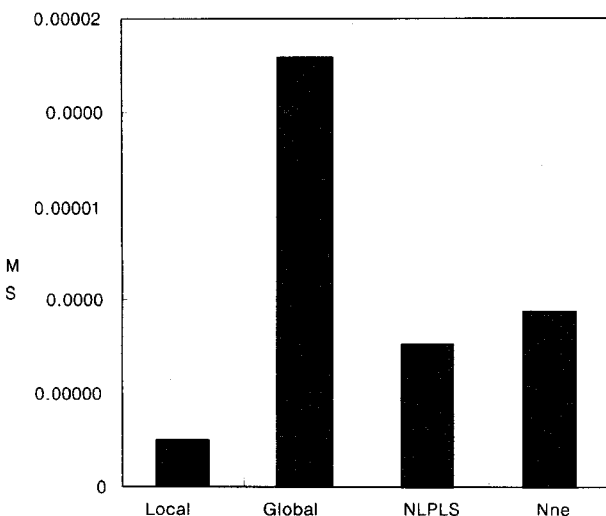


Fig. 14. Comparison of estimation.

조성의 변화로 인해 조업 영역이 넓어서 조업 영역이 여러 개로 나뉘는 때, 각각에 해당되는 국부적인 조성 추정 소프트웨어를 PLS방법으로써 설계하고 global PLS, 신경회로망을 기반으로 하는 비선형 PLS, 신경회로망 등을 이용한 소프트웨어와 모델의 성능 및 예측 성능을 비교하였다.

Linkage based clustering방법은 비선형성이 강하고 covariance 구조가 원형이 아닌 경우 효과적으로 조업 영역을 구분할 수 있음을 보였다. 또한 linkage based clustering 방법을 통한 데이터 분류 과정에서 데이터 기반 소프트웨어의 설계에 중요한 모델링 전의 특이한 데이터가 제거되므로 신뢰성있는 소프트웨어를 설계할 수 있었다.

PLS는 차원이 크고 상관성이 높은 데이터를 효과적으로 다루어주는 통계적인 모델링 방법으로써 측정 잡음이나 센서의 고장에 강건한 것으로 알려져 있다.

본 연구에서 제안하는 소프트웨어는 비선형성이 강한 산업체 증류 공정에 국부 PLS모델을 이용해 설계된 것으로 뛰어난 조성 추정 능력을 보인다.

감 사

본 연구를 위해 포항공대 공정산업의 지능자동화연구센터를 통해 재정적 지원을 해주신 한국과학재단에 감사를 드립니다.

사용기호

- X_{ij} : data matrix with i rows and j columns
- \bar{X}_j : mean value of j th column
- σ_j : standard deviation of j th column
- K : number of nearest neighbor points to be a linkage
- C : number of clusters
- a : percent of outliers within total data
- I : index of total data
- J : index of data clustered
- L : index of data nonclassified
- NN_{ik} : k th nearest neighbor point in i th cluster
- S_c : seed point to begin the clustering in c th cluster
- CP_{cj} : j th data clustered in a c th cluster
- NP_l : l th data nonclassified of total data
- a : number of PC
- N : number of observations
- \hat{y}_i : estimated composition at i th observation
- \bar{y} : mean value of observations
- y_i : measured real value at i th observation

참고문헌

1. Tham, M. T., Montague, G. A., Morris, A. J. and Lant, P. A.: *J. Proc. Cont.*, 1, 3(1991).
2. Shin, J., Lee, M. and Park, S.: *Korean. J. Chem. Eng.*, 15, 6(1998).
3. Znar, A.: "Nonlinear Time Series Models for Multivariate Dynamic Processes," in web site of Ali Znar(1996).
4. Kresta, J. V., Marlin, T. E. and MacGregor, J. F.: *Computers chem. Eng.*, 18, 7(1994).
5. Fujii, H., Lakshminarayanan, S. and Shah, S. L.: "Application of the PLS Technique to the Estimation of Distillation Tower Top Composition," IFAC ADCHEM, 529(1997).

6. Mejedell, T. and Skogestad, S.: "Output Estimation Using Multiple Secondary Measurements: High-Purity Distillation," **39**, 10(1993).
7. Geladi, P. and Kowalski, B. R.: *Analytica Chimica Acta*, **185**, 1 (1986).
8. Kresta, J. V., Marlin, T. E. and MacGregor, J. F.: *Computers chem. Eng.*, **18**(7), 597(1994).
9. Martens, H. and Naes, T.: "Multivariate Calibration," John Wiley & Sons(1989).
10. Qin, S. J. and McAvoy, T. J.: *Computers chem. Engng.*, **16**(4), 379 (1992).
11. Willett, P.: "Similarity and Clustering in Chemical Information Systems," John Wiley & Sons(1987).
12. Duda, R. O. and Hart, P. E.: "Pattern Classification and Scene Analysis," John Wiley & Sons(1973).