총계오차판별 및 데이터 보정 개요

포항공대 화학공학과 김 정 환

서울대학교 응용화학부 한 종 훈

본 강좌에서는 데이터 신뢰성 확보를 위한 중요기술인 총계오차판별 (Gross error detection) 및 데이터보정 (data reconciliation) 기술에 관하여 총 3 회에 걸쳐 살펴보도록 하겠다. 본 회에서는 데이터보정의 기본개념에 대하여 소개하고, 다음 회에서는 다양한 총계오차판별기법 및 연구흐름 소개를, 마지막 회에서는 현장적용 사례와 시스템 구성에 관하여 소개하도록 하겠다.

1. 총계오차판별 및 데이터보정 개요

대부분의 화학공정에는 수백 개 혹은 수천 개의 센서들이 존재하며, 유량, 온도, 압력, 레벨 등 다양한 정보를 측정한다. 과거의 데이터 수동기입과 비교하면 데이터수집의 자동화 자체만으로도 데이터 정확성을 많이 향상시켰지만 늘어난 정보량만큼이나 데이터 처리 정확성 향상의 필요성 또한 높아지고 있다.

측정값은 불가피하게 오차를 포함할 수 밖에 없는데, 오차의 원인들로서는 측정오차, 측정신호의 처리 및 전송오차 등이 있으며, 이런 오차 (실제값과 측정값의 차이)는 총계오차(Gross error)와 확률오차(random error)의 합으로 표현될 수 있다. 확률오차란 그 크기나 방향이 전혀 예측이 안 되는 오차로서 동일한 시스템에 대하여 동일한 조건의 측정을 하여도 확률오차의 영향에 따라서 측정값이 달라지게 되며, 확률오차의 특성에 대해서는 확률분포로서 표현을 하게 되며 일반적으로 그 값이 작다. 확률오차는 주로 측정값을 전기적 신호로 바꿀 때 발생하는 노이즈나 전원공급의 불규칙성, 아날로그 신호 필터링 등에서 발생하는 경우가 많다. 이와 반대로, 특정한 방향성을 갖고 반복적으로 되풀이되는 오차를 총계오차라고 한다. 총계오차의 발생원인은 측정장치의 고장이나 캘리브레이션

오차 (계측장비나 기타 장비의 측정 기준점 및 스케일이 부정확하여 발생하는 오차), 센서고장 등이다.

총계오차 및 확률오차가 제거되지 않은 데이터를 프로세스에서 그대로 사용하게 되면 프로세스 성능 모니터링의 정확성 저하, 원가관리 결과의 신뢰성 저하, 컨트롤 시스템의 성능저하 및 최적화의 경우 그 결과를 전혀 신뢰할 수 없게 된다. 따라서, 데이터보정 및 총계오차판별은 실시간 최적화를 위한 매우 기초적이며 중요한 작업이 된다.

데이터의 정확성 향상을 위해서 총계오차판별을 통한 총계오차제거를 먼저 수행한 후, 데이터보정을 통하여 확률오차를 제거하게 된다. 데이터보정기법이 데이터 정확성 향상을 목표로 하는 다른 필터링 기법들과 구분되는 것은 확률오차의 감소를 통하여 시스템에 존재하는 제약조건들을 만족시키는 값을 찾아내도록 한다는 점이다. 데이터 보정을 통하여 확률오차를 제거함으로써 데이터의 정확성을 높일 수 있으며, 이 데이터들은 시스템에 존재하는 제약조건들을 모두 만족시키는 값을 얻도록 한다는 점이 중요한 포인트이다.

2. 데이터 보정 문제 정형화

우선, 데이터보정을 통해서 데이터의 신뢰성을 향상시키기 위해서는 대상시스템에 Redundancy 가 존재해야만 한다. 일반적으로 센서 설치비용은 매우 비싸기 때문에 모든 지점에 센서를 설치한다는 것은 불가능하고 꼭 필요한 부분의 값을 측정하게 된다. 따라서, 꼭 필요한 곳에 측정센서를 설치하도록 하는 Sensor Network 결정도 매우 중요한 분야이다. 일반적으로 데이터보정 문제는 다음과 같이 표현된다.

$$\underset{X_i,U_j}{Min} \sum_{i=1}^{n} w_i (y_i - x_i)^2 \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (1)$$

subject to

$$g_k(x_i, u_j) = 0, \quad k = 1,..,m \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (2)$$

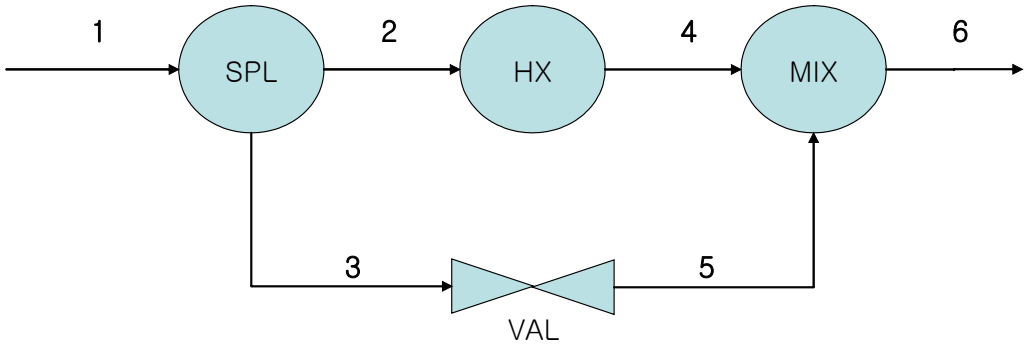
위 식에서, x_i 는 i 측정변수에 대한 보정값이며, u_j 는 미측정변수에 대한 예측값이다. w_i 는 센서의 정확성을 나타내기 위한 가중치이며, y_i 는 i 측정변수에 대한 측정값이다. 즉, 목적함수는 현재 측정되고 있는 각 센서의 조정값에 대한 가중총합을 최소화하면서 시스템 내의 제약조건들을 만족시키는 것이다. 여기서, 가중치는 각 측정장치의 정확도를 반영하게 된다. 즉, 현재 장치가 매우 정확한 장치라면 큰 가중치가 곱해짐으로써 가중총합을 최소화시키는 최적화 목적함수이므로 조정값을 조금만 변화시키고 대신 정확도가 낮아서 가중치가 낮은 항목에 대해서 더 많은 조정을 하도록 결정된다. 일반적으로, 이 가중치값은 각 측정치의 오차의 분포를 조사하여 분산의 역수를 취한다.

제약조건들은 주로 물질수지식과 에너지 수지식이다. 한가지 주의해야 할 점으로서는, 물질수지식이나 에너지수지식 이외에 경험식이나 소프트센서수식 등을 제약조건으로 첨가하는 것은 바람직스럽지 않다는 점이다. 왜냐하면, 이런 제약조건들은 기껏해야 그 변수에 대한 근사식이기 때문에 또 다른 에러의 근원만 될 뿐이기 때문이다. 따라서, 가장 정확한 물리현상인 물질수지식과 에너지수지식을 이용하는 것이 바람직하다.

그런데, 여기서 한 가지 주의할 점은 물질수지식과 에너지수지식을 제약조건으로 활용할 경우에 가정은 시스템에서 물질이나 에너지가 손실되거나 추가되지 않는다는 가정이 존재한다. 예를 들어, 제대로 보온장치가 안 되어 있어서 열손실이 발생하는 경우에는 에너지수지식을 그대로 적용하기 보다는 에너지 손실항목을 추가하여 손실부분을 반영하여 제약조건을 구성해야 한다. 간단한 예를 통해서 데이터보정을 통한 데이터 신뢰성 향상효과를 확인해 보도록 하자.

3. 데이터 보정 적용 예

<그림 1>과 같은 Bypass 가 존재하는 간단한 열교환시스템을 대상으로 데이터보정의 수행에 관하여 알아보도록 하겠다. 현재 센서는 모두 6 개가 존재하며 본 예에서는 물질수지식에 관한 제약조건만을 고려한다.



<그림 1> Heat Exchanger System

3.1. 시스템 내 모든 변수가 측정 가능한 경우 (Observable & Redundant Case)

먼저, 가장 간단한 경우로서 측정변수 1~6 이 모두 측정되는 경우이다. 이 경우 각 장치의 input 과 output 의 밸런스를 만족하는 제약조건이 모두 4 개가 존재하게 된다. 목적함수 및 제약조건은 다음과 같다.

$$\underset{X_i}{Min} \sum_{i=1}^{6} (y_i - x_i)^2 \quad \text{..... (3)}$$

subject to

$$x_1 - x_2 - x_3 = 0 \quad \text{..... (4)}$$

$$x_2 - x_4 = 0 \quad \text{..... (5)}$$

$$x_3 - x_5 = 0 \quad \text{..... (6)}$$

$$x_4 + x_5 - x_6 = 0 \quad \text{..... (7)}$$

이 최적화 문제를 푼 결과는 표 1-1 과 같다.

<표 1.1> 모든 변수가 측정가능한 경우의 데이터보정 수행 결과

스트림	참값	측정값	보정값
1	100	101.91	100.22
2	64	64.45	64.50
3	36	34.65	35.72
4	64	64.20	64.50
5	36	36.44	34.70
6	100	98.88	100.22

데이터 보정을 통하여 각 장치의 입,출력간 밸런스가 맞지 않던 유량들이 모두 밸런스를 맞추도록 조정되었음을 알 수 있다. <표 1.1>에서 참값이라고 하는 것은 참값이라고 가정한 값이다. 실제로 참값이 얼마인지는 정확히 알 수 없다.

3.2. 시스템 내 미측정 변수가 존재하는 경우

시스템 내에 미측정변수가 존재하는 경우에는 어떤 변수가 미측정변수이며 그 변수와 관련있는 제약조건들이 어떤 제약조건인가에 따라서 데이터보정 결과는 많이 달라진다. 첫번째 케이스는, <그림 1>의 시스템에서 3,4 번 스트림이 측정되지 않고, 나머지 스트림의 유량은 모두 측정되는 경우이다. 이 경우, 다음과 같은 데이터보정 문제 정형화가 된다.

$$\underset{x_1,x_2,x_5,x_6}{Min}\ (y_1-x_1)^2+(y_2-x_2)^2+(y_5-x_5)^2+(y_6-x_6)^2 \quad \text{..... (8)}$$

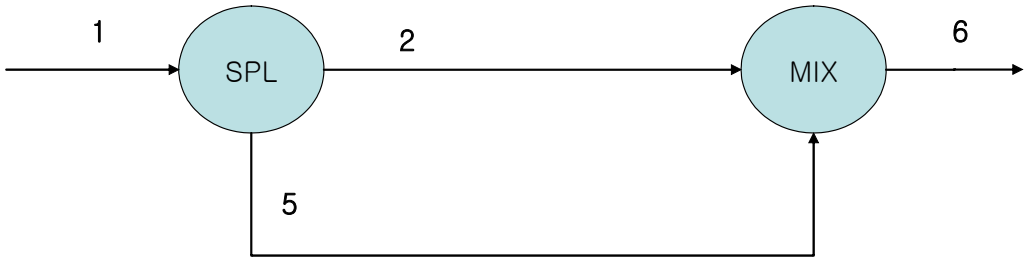
subject to

$$x_1-x_2-x_5=0 \quad \text{..... (9)}$$

$$x_2+x_5-x_6=0 \quad \text{..... (10)}$$

제약조건 (9)와 (10)은 다른 제약조건식을 이용하여 측정되는 변수들만을 이용한 새로운 제약조건식을 구성한 것이다. 식(8)~(10)을 이용하여 최적화 문제를

풀면 <표 1.2>와 같은 결과를 얻게 된다. 본 최적화 문제의 경우 모든 변수들이 측정되는 변수들로만 구성이 되어 에러의 합을 최소화시키는 보정값을 얻은 후에 제약조건들식을 이용하여 나머지 미측정변수에 대한 예측이 가능한 경우이다. 즉, <그림 1>의 시스템대신 <그림 2>의 시스템을 푸는 문제가 되며, 이 문제를 풀어서 결정된 스트림값 1,2,5,6 을 이용하여 미측정 변수인 3,4 를 제약조건식으로부터 구하게 되는 것이다. 이와 같이 데이터보정을 통하여 결정된 값을 이용하여 미측정값을 예측하는 것을 *coaptation* 이라고 한다.



<그림 2> Reduced System

<표 1.2> 미측정값이 존재하는 시스템의 케이스별 데이터보정결과

스트림	Case1 스트림 3,4 미측정	Case2 스트림 3,4,5,6 미측정	Case3 스트림 2,3,4,5 미측정
1	100.49	101.91	100.39
2	64.25	64.45	-
3	36.24	37.46	-
4	64.25	64.45	-
5	36.24	37.46	-
6	100.49	101.91	100.39

두번째 케이스는 스트림 3,4,5,6 이 미측정되고 있는 경우이다. 이 경우에도 앞에서 설명된 내용대로 목적함수와 제약조건을 구성할 수 있는데, 케이스 1 처럼 제약조건이 측정변수들로만 구성된 제약조건을 만들수가 없어서 케이스 1 처럼 측정변수가 보정될 수가 없는 상황이다. 즉, 측정되는 변수 1 과 2 는 제약조건들에 의해서 보정이 될 수 없기 때문에 목적함수를 가장 최소화시키는 값은 현재의 측정값으로 결정되는 케이스이다. 스트림 1 과 2 의 값은 이렇게 조정되지 않으며 나머지 값들은 제약조건식들에 의해서 결정된다. 이 경우, 측정변수 1 과 2 는 redundant 하지 않다고 말한다. 또한, 미측정변수 3,4,5,6 의 경우는 측정값과 제약조건들에 의해서 유일한 값으로 결정될 수 있기 때문에 observable 하다고 말한다.

세번째 케이스는 시스템 중에서 처음 입력부분과 마지막 출력부분의 값만 측정되며 가운데 부분의 값은 전혀 측정되지 않는 케이스이다. (즉 스트림 2,3,4,5 미측정 케이스) 이 경우 데이터보정 문제 구성은 다음과 같다.

$$\underset{x_1,x_6}{Min}(y_1 - x_1)^2 + (y_6 - x_6)^2 \quad \text{..... (11)}$$

subject to

$$x_1 - x_6 = 0 \quad \text{..... (12)}$$

이 경우, 목적함수와 제약조건은 측정변수들로만 구성이 가능하다. 따라서, 측정변수인 스트림 1,6 의 경우 redundant 하며 미측정변수들의 경우 다양한 해가 존재하므로 유일한 해로서 결정할 수가 없다. 따라서 미측정변수 2,3,4,5 는 unobservable 하다. 즉, <표 1.2>에 보이는 바와 같이 스트림 2,3,4,5 에 대하여는 값을 결정할 수가 없다.

이와 같이 측정변수 및 이와 관련된 제약조건에 의하여 미측정변수값을 예측하는 것이 가능한 경우도 있고 불가능한 경우도 존재하게 된다. 따라서, 적절한 센서의 위치를 결정하는 것이 중요함을 알 수 있다. 참고로, 데이터 보정에서 사용하는 중요한 개념인 Observability 와 Redundancy 의 정의는 다음과 같다.

Observability: A variable is said to be observable if it can be estimated by using measurements and steady-state process constraints.

Redundancy: A measured variable is said to be redundant if it is observable even when its measurements is removed.

4. 데이터 보정과 총계오차판별의 활용분야

- ✓ 프로세스 수율 (Yield) 분석

  데이터보정을 통하여 대상 프로세스의 수율에 대한 분석이 가능하며 원가관리가 가능하다. 특히나, 서로 이해관계가 상충하는 부서의 의견충돌을 해결하는 것이 가능하다. (예를 들어, 공정내에서 스팀을 생산하는 부서와 활용하는 부서간의 생산량과 소비량에 대한 불일치 해결)
- ✓ 최적화 및 시뮬레이션을 위한 모델 튜닝

  최적화 및 시뮬레이션을 구성하는 여러 파라미터들은 데이터를 기반으로 산출한 것이 많은데, 이 파라미터들이 부정확할 경우 잘못된 시뮬레이션 결과를 얻게 되며 최적화 결과 또한 신뢰하기 어렵다. 따라서, 데이터 보정은 시뮬레이션이나 최적화를 정확히 수행하기 위해서는 필수적인 요소이다.
- ✓ 프로세스 장치에 대한 수리계획 수립

  데이터보정을 통하여 프로세스 내 주요장치의 KPI 값을 모니터링함으로써 주요장치의 성능 저하시에 신속하게 수리함으로써 공정의 전체적인 성능을 높게 유지하는 것이 가능하다.
- ✓ 측정장치의 고장 감지

  총계오차판별기법을 적용함으로써 데이터보정의 정확성을 향상시킬 뿐만 아니라 공정 내 센서의 이상유무를 판별하여 켈리브레이션이나 센서교체 등과 같은 작업을 수행할 수 있다.

다음 회에서는 총계오차판별을 위한 다양한 기법과 데이터보정 및 총계오차판별 연구흐름에 관하여 살펴보도록 하겠다.