# 데이터 전처리

## 화공 데이터의 전처리

화공 산업의 현상을 정교하게 분석하기 위해선 데이터의 수준을 향상시킬 필요가 있다. 특히 실제 산업 현장에서 수집된 다양하고 방대한 데이터는 데이터 자체의 특이적 변형, 데이터 획득 과정에서의 오류, 수집과정에서의 손실 및 변형 등 데이터 처리 과정에서 다양한 원인으로 데이터의 양과 질이 왜곡될 수 있다. 특히. 문제해결의 과정과 결과에 큰 영향을 미치는 다양한 데이터의 오류는 인공지능을 이용한 문제해결 전 단계에서 반드시 수정 및 보정이 필요하다. 따라서 적합한 데이터 전처리 과정은 인공지능 모델 개발에 있어 필수적이고 데이터 가치를 보존할 수 있는 방법이다. 본 장에서는 데이터의 가치를 확보하면서 데이터의 오류 등을 관리할 수 있는 데이터 정제, 이상치 제거, 그리고 변수 규명 방법 등 이른바 데이터 전처리 기술에 관하여 학습한다.

### 데이터 정제 (Data cleaning)

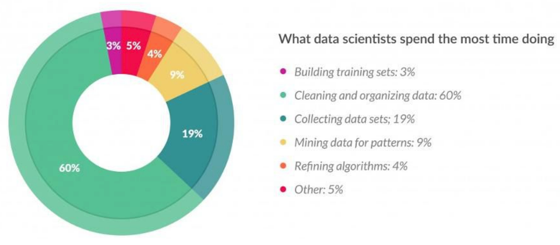


그림 1. 데이터 분석 과정별 소모 시간.

데이터 정제 방법은 관련성 높은 데이터들을 추출하고 데이터 오류를 파악 · 제거하는 방법으로 데이터의 신뢰도를 향상시킨다. 데이터 정제는 데이터 분석 프로젝트를 수행하면서 가장 많은 시간과 노력을 요구한다. 이는 수집된 원시 데이터가 보통 많은 결함을 가지고 있어 분석 · 해석에 바로 사용되기 어려운 특징 때문이다. 데이터 정제 기술은 세가지 기법: 변환 (Transformation), 분할 (Parsing), 그리고 보강 (Enhancement)을 포함한다. 각 기법이 의미하는 바는 아래와 같다.

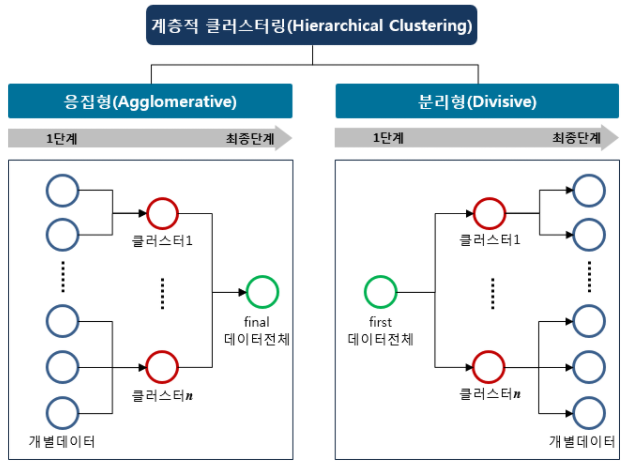
- 변환: 다른 규격 · 형태의 데이터들을 일관된 형태로 수정

- 분할: 데이터 일관성 향상을 위해 작은 단위의 데이터로 분할

- 보강: 데이터 표준화 같은 방법을 통해 데이터의 가치를 향상

데이터 정제 활동을 통해 세가지 데이터 정제 기법을 수행하면 데이터 품질을 확보할 수 있다. 예를 들면, 데이터 수집단계에서의 오류 원인 파악과 정제 대상에 따른 적절한 데이터 정제 기법을 통해 데이터의 품질을 확보할 수 있다. 데이터 정제 활동은 데이터 세분화, 결측치 처리, 그리고 이상치 처리로 세분화될 수 있다.

* 데이터 세분화 (Data segmentation)



* 그림 2. 계층적 데이터 세분화의 예시 (좌) 응집형 세분화, (우) 분리형 세분화.

데이터 세분화는 특정 기준을 제시하여 데이터를 유의미한 최소 단위로 나누고 유사한 데이터들을 그룹화하는 데이터 정제 활동이다. 또한, 단일화된 통계에서 나타나지 않는 특징이 가시적으로 보이도록 돕는 작업이다. 흔히 군집을 사전에 정하고 데이터 군집을 단계적으로 산출하는 계층적 방법이 많이 쓰인다. 이는 개별 데이터로부터 군집 기준을 정하는 응집분석법과 전체 집단에서 군집 기준을 정하는 분리분석법이 존재한다.

* 결측치 처리

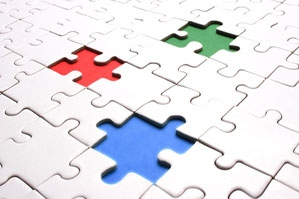


그림 3. 빠진 퍼즐과 같은 데이터 결측치.

데이터에서 결측치란, 데이터 수집 · 이동 · 저장 중 데이터의 손실 혹은 오류로 인해 누락되거나 변형된 값을 의미한다. 결측의 원인 파악은 결측값의 판단과 식별에 매우 중요한 요소로, 데이터 결측값 처리 방법을 결정한다. 데이터 결측값은 원인에 따라 다양한 형태로 표현된다.

- NULL: 데이터 자료가 아예 존재하지 않는 경우.

- NA (Not available): 이동 · 저장 중 오류로 사용할 수 없는 데이터.

- NaN (Not a number): 수치적으로 정의되지 않은 경우.

- Inf (Infinite): 무한을 의미하며 잘못된 수치적 계산에서 발생.

결측치를 효과적으로 처리하기 위한 방법으론 단순 대치법 (Single imputation)과 다중 대치법 (Multiple imputation)이 활용된다. 단순 대치법은 자료 무시 (완전 분석법), 평균치 사용 (평균 대치법), 그리고 확률값 부여 (단순 확률 대치법) 방법을 통해 적절할 값으로 결측치를 대체한다. 다중 대치법은 자료들을 예측된 분포를 사용하여 결측값을 대체하는 방법이다.

* 이상치 (Outlier) 처리

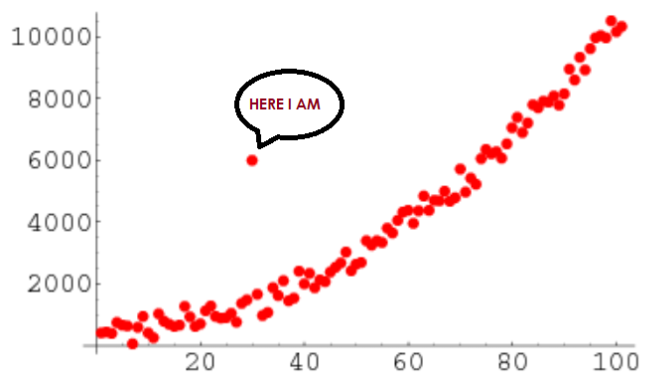


그림 4. 이상치의 예시.

데이터 이상치는 수집된 데이터의 범위를 크게 벗어난 소수의 자료들을 의미한다. 보통, 데이터 센서와 같은 장치를 통한 수집이나 데이터 이동시 처리 오류들로 인해 데이터 값이 잘못될 수 있다. 오류의 발생 원인에 따라 입력 오류, 측정 오류, 실험 오류, 기입 오류, 그리고 추출 오류가 존재한다. 데이터 분석 시 이상치는 현상의 경향을 파악하는데 치명적이게 작용하므로, 이상치를 검출 및 처리하는 것이 중요하다. 다음과 같이 다양한 데이터 이상값 검출 방법은 존재한다.

- 통계 기법: 데이터의 통계적 지표 (평균, 분산 등)을 활용한 검출.

- 시각화 기법: 히스토그램과 같은 분포 그래프를 이용한 이상값 검출.

- 머신러닝 기법: 비지도 학습 중 군집화 알고리즘을 사용한 검출.

- 마할라노비스 거리 (Mahalanobis distance): 데이터 거리 측도를 사용한 검출.

- LOF (Local outlier factor): 데이터들의 상대적인 밀도 비교를 통한 검출.

- iForest (Isolation forest): 의사결정나무를 활용한 특정 데이터 식별 요구 단계 횟수로 검출.

또한, 다음과 같이 다양한 이상치 처리 방법론을 사용하여 해당 자료를 수정할 수 있다.

- 삭제: 이상치로 판단되는 데이터를 데이터 베이스로부터 제거.

- 대체: 데이터 분포를 고려하여 평균이나 중앙값 등 통계값으로 대체.

- 변환: 극단적인 데이터 (지수적 상승)의 경우, 자연로그 데이터 변환.

- 분류: 이상치가 많은 경우, 해당 데이터를 분리시켜 따로 분석 실행.

* 정규화/표준화

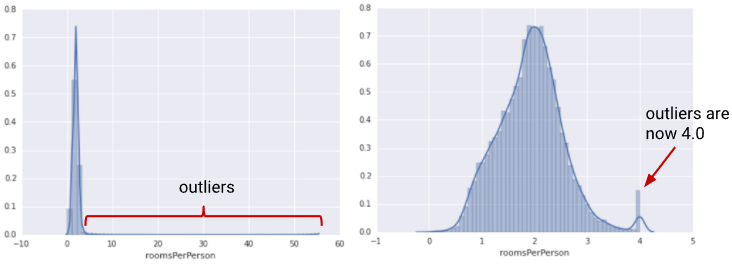


그림 5. 정규화/표준화를 통한 데이터 품질 향상 예시.

화공 산업의 데이터는 서로 다른 단위 (℃, bar, wt%, km 등)와 크기를 갖는다. 인공지능은 학습 데이터의 변위를 기반으로 데이터의 영향력을 이해하므로, 높은 수준의 인공지능 모델 개발을 위해서 데이터 특성 적절히 변환시키는 과정이 필요하다. 대표적인 방법으로 정규화 (Normalization)과 표준화 (Standardization) 이 있으며 다음과 같은 공식으로 표현된다.

정규화:

표준화:

*X*는 자료의 값을 의미하고, *Xmin*과 *Xmax*는 각각 자료의 최대 및 최솟값을 의미한다. *σ*와 *μ*는 자료의 표준편차와 평균을 의미한다. 데이터에 따라 좋은 방법이 다르기 때문에 두가지 방법 모두 활용하여 모델 학습에 이용하는 것이 적절하다.

### 데이터 변수 규명

화공 산업에서는 장치마다 다양한 공정 변수 (물질 유량, 온도, 압력 등)에 의해 조작된다. 공정의 기술성 · 경제성 · 환경성 등은 이러한 공정 변수를 독립 변수로 결정되는 종속 변수다. 하지만, 모든 독립 변수를 고려하는 것은 현상을 지나치게 복잡하게 정의하는 것이며 해석에 많은 데이터와 노력을 필요로 한다. 공정 데이터 해석에는 크게 변수 선택과 차원 축소 절차가 있다.

* 변수 선택

변수 선택은 종속 변수 해석에 가장 중요하고 관련성 높은 독립 변수를 추출하는 과정이다. 또한, 구체적인 변수 명칭, 인과 관계, 변수 속성 등 다양한 변수 특성과 유형을 파악하여 사용자가 해석에 용이하도록 데이터 구성을 단순화하는 기능을 한다. 변수 선택을 위한 방법에는 다음과 같은 기법이 자주 수행된다.

- 필터 기법 (Filter method): 변수들 사이의 상관관계 파악을 통한 통계적인 측정 방법.

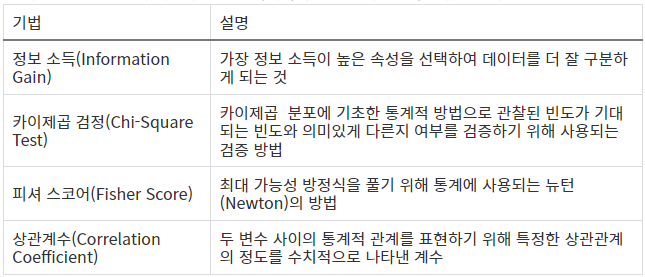


그림 6. 필터 기법의 예시.

- 래퍼 기법 (Wrapper method): 모델 개발 시 예측 정확도에 큰 영향을 미치는 집합을 선택하는 기법.



그림 7. 래퍼 기법의 예시.

- 임베디드 기법 (Embedded method): 모델 개발 시 예측 정확도에 기여하는 변수를 선택하는 기법.

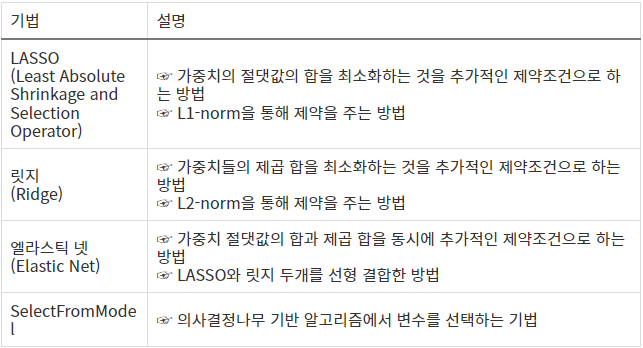


그림 8. 임베디드 기법의 예시.

* 차원 축소 (Dimension reduction)

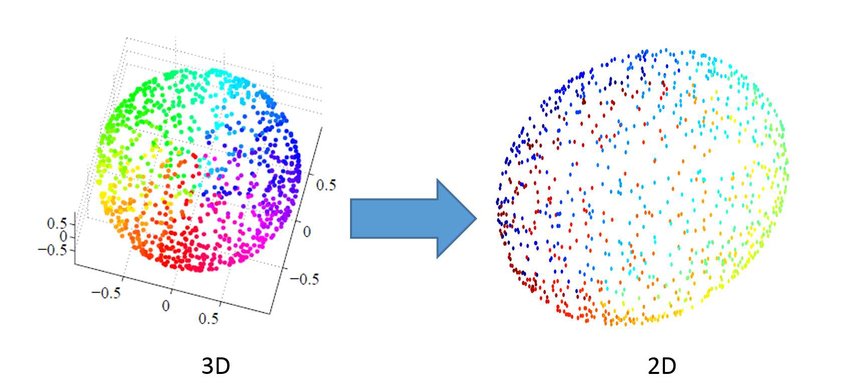


그림 9. 차원 축소 방법론을 통한 데이터 압축 예시.

차원 축소는 최대한 기존 정보를 유지하면서 고차원 데이터를 낮은 차원으로 투영하는 방법이다. 일반적으로 차원이 증가하면 데이터의 밀도가 낮아져서 더욱 많은 데이터가 필요한 문제가 발생한다. 이를 차원의 저주 (Curse of dimensionality)라고 하며, 불필요한 차원을 제거하는 방식으로 해결할 수 있다. 이런 경우, 차원 축소 방법론은 데이터의 변수 개수를 효과적으로 줄여 시각화 및 분석에 유용하게 쓰인다. 이를테면 차원 축소는 데이터를 압축하고 잠재적인 정보를 추출하여 인공지능 모델 개발 시 학습 시간을 효과적으로 낮추고 예측 모델의 신뢰성을 향상시킬 수 있다. 독립 변수의 종속 변수에 대한 기여도와 특성 · 관계성 파악에 쓰이며, 보통 다른 분석 (예측 모델 개발 등)의 선행적 단계에 시행된다.

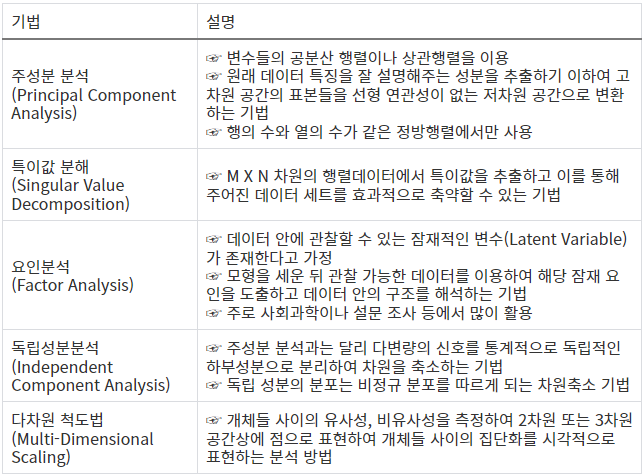


그림 10. 차원 축소의 예시.

## 학습 결과

* 학습 내용

화공 데이터 전처리를 위한 데이터 정제와 변수 규명 방법론.

* 학습 결과 확인하기

화공 산업에 효과적인 빅 데이터 활용을 위해, 주요 데이터 정제 과정과 변수 규명 방법 익히기.

* 학습 결과 응용하기

본 장의 학습내용에 기반해 화공 산업에서 발생하는 데이터의 전처리 역량을 확보함으로, 실제 화공 산업 현장의 데이터 분석 및 모델 개발을 위한 데이터 처리에 응용.