데이터 전처리 주요 기법

- 데이터 정제(Data Cleansing)

  없는 데이터(missing values)는 채우고, 잡음(noisy data)는 제거하며, 모순된 데이터(inconsistent data)는 정합성이 맞는 데이터로 교정하는 작업
- 데이터 통합(Data Integration)

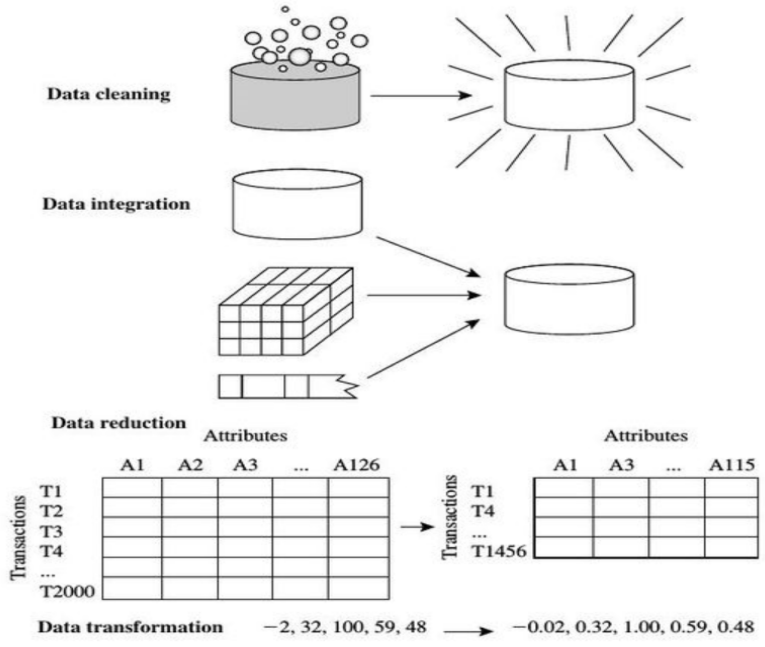
  여러 개의 데이터베이스(databases), 데이터큐브(data cubes), 또는 파일(files)을 통합하는 작업
- 데이터 축소(Data Reduction)

  샘플링(sampling) 등을 통해 데이터 볼륨(volume)을 줄이거나 분석대상 속성(차원)을 줄이는 작업
- 데이터 변환(Data Transformation)

  데이터 정규화(normalization) 또는 집단화(aggregation) 하는 작업
- 데이터 이산화(Data discretization)

  데이터 축소(data reduction)의 일종으로 연속적인 수치 데이터에 대한 구간화 작업 (예, 실제 나이를 10대, 20대, 30대 등으로 변환)

데이터 전처리 주요 기법



* 출처 : Jiawei Han and Micheline Kamber. Data Mining: Concepts & Techniques

[그림 1.6] 데이터 전처리 주요 기법의 개념 (concept)

학습의 방향

- 데이터 전처리 유형
  - 실무에 있는 데이터는 매우 다양한 형태로 존재하기 때문에 이를 특정 분석 목적에 맞게 가공하는 일은 사안마다 다름
  - 데이터 전처리를 개념적인 몇가지 분류 내로 국한시켜서는 분석 대상 데이터를 만들어내기 힘듬
  - 데이터 전처리는 개념적·이론적으로 제시되는 범위보다 훨씬 넓고 포괄적
- 학습의 방향
  - 모든 데이터 전처리 유형을 다루는 것은 애초부터 불가능 → 이론적으로도 제시되어 있으면서 실무에 자주 등장하는 전처리 유형을 중심으로 학습
  - 이론적 이해 + 실무 사례 적용을 통한 활용능력 배양
- 활용 도구(tools)
  - Python : 플랫폼 독립적이며 인터프리터식, 객체지향적, 동적 타이핑(dynamically typed) 대화형 언어로서 데이터프레임, 기계학습 등의 데이터분석을 위한 다양한 함수를 제공하여 데이터전처리에 적합한 도구
  - Oracle : Python의 SQL은 오라클의 SQL에서 제공하는 윈도우(window) 함수와 같은 데이터의 집합적 처리를 위한 강력한 기능은 제공되지 않은 부분이 있어서, 일부 사례에서는 오라클 SQL을 통해 해답을 제시