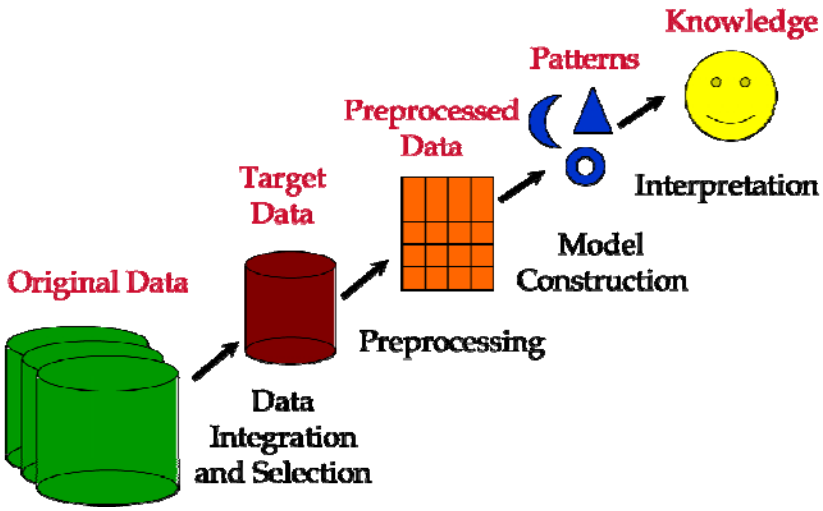
Ch.1 데이터전처리 개요

데이터전처리 개요

- 데이터 전처리 정의

  데이터 분석 작업을 하기 전에 데이터를 분석하기 좋은 형태로 만드는 과정을 총칭하는 개념



* 출처 : Ramakrishnan and Gehrke. Database Management Systems, 3rd Edition

[그림 1.1] 데이터마이닝 과정에서의 데이터 전처리

데이터전처리 개요

- 데이터 전처리가 필요한 이유
  - 실무 데이터는 분석 기법을 바로 적용하기 힘든 형태
  - 비어있음(missing value), 잡음(noise), 적합하지 않은 데이터구조
  - 낮은 품질의 데이터로는 좋은 분석결과를 얻기 힘듬
- 데이터 품질 저하의 원인
  - 불완전(incomplete)
    - 데이터가 비어 있어 있는 경우로 DB 테이블의 속성값이 NULL인 경우
  - 잡음(noisy)
    - 데이터에 오류(error)가 포함된 경우. 예) 나이 = -20
  - 모순된(inconsistent)
    - 데이터 간의 정합성(일관성)이 없는 경우. 예) 성별은 남자인데, 주민번호 뒷 7자리 중 첫 번째 자리가 2인 경우
- 고품질 데이터라 하더라도 전처리 필요
  - 실무에서 존재하는 데이터의 구조적 형태(format)가 분석목적이나 분석기법에 적합한 경우가 드물기 때문