데이터마이닝(DataMining)

Chapter 1. 데이터마이닝의 개념

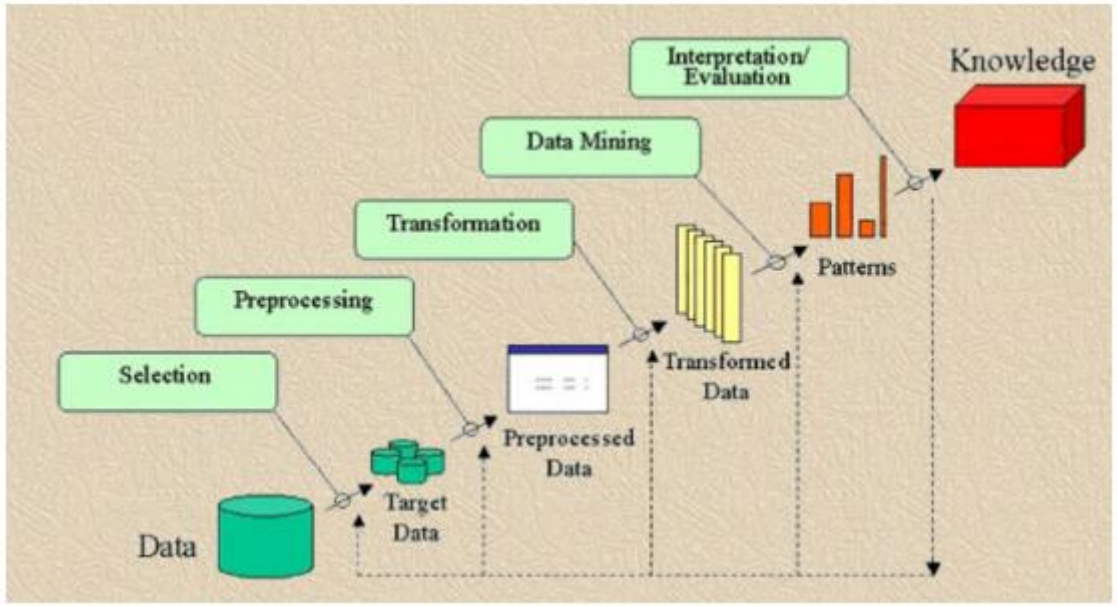
데이터마이닝이란?

- 데이터마이닝 (data mining)은 대량의 데이터로부터 규칙이나 패턴을 찾아내는 과정으로, 통계학, 데이터베이스, 기계학습, 인공지능의 영역에서 발전된 다양한 기법들을 포함

- 데이터마이닝의 목적은 데이터셋으로부터 정보를 추출하고, 이를 추후 사용을 위해 이해할 수 있는 구조로 변환하는 것

- 데이터마이닝은 분석 단계 이외에도 데이터베이스와 데이터 관리, 데이터 전처리, 모형과 추론 고려사항, 흥미도, 복잡성, 발견된 구조의 사후처리, 시각화 및 온라인 업데이트 등을 포함

- 데이터마이닝은 KDD(knowledge-discovery in databases, 데이터베이스 속의 지식발견) 과정 또는 KDD 과정의 분석 단계로 이해될 수 있음



- 데이터마이닝의 적용분야는 매우 다양

- 기업에서는 표적마케팅, 고객세분화, 고객성향 분석 등에 활용하고 있으며, 금융 분야에서는 신용평가, 거래사기 적발 등에 활용

- 제조업에서의 품질관리, 의학분야에서의 유전자 분석, 지구과학 및 천문분야에서의 방대한 자료처리에 활용

- 텍스트마이닝을 통한 정보검색과 음성과 영상 등의 멀티미디어 자료의 분석에도 활용

- 빅데이터 분석에서도 데이터마이닝은 핵심적인 역할을 담당

지도학습과 비지도학습

- 예측모형은 결과값이 알려진 다변량 자료를 이용하여 모형을 구축하고, 이를 통해 새로운 자료 에 대해 결과값에 대한 예측 또는 분류를 수행하는 방법

- 결과값이 범주형인 경우에는 새로운 자료에 대한 분류(classification)가 주목적이며, 결과값이 연속형인 경우에는 예측(prediction)이 주목적

- 예측과 분류는 유사한 의미로 사용되며 통칭하여 예측모형 부르기도 함

- 대표적인 예측모형으로는 로지스틱 회귀, 의사결정나무, 판별분석, 인접이웃분류, 베이즈분류, 신경망, 서포트벡터머신과 이들 예측모형(분류기)들을 결합한 앙상블 모형 등

- 기계학습 분야에서는, 결과값이 알려진 상황에서의 학습모형인, 예측모형을 지도학습 (supervised learning)이라 부른다. 예측모형은 목표마케팅, 성과예측, 의학진단, 사기검출, 제 조 등 다양한 분야에 이용

- 예측모형과는 달리 별도의 결과값을 요구하지 않는 자료에 대한 분석을 비지도학습 (unsupervised learning)

- 군집분석은 데이터의 개체들 간의 유사성에 기반하여 전체 개체를 몇 개의 군집으로 나누는 방법으로 사용. 모형 구축시에 결과값이 주어져 있지 않음으로 오차(또는 보상 신호) 의 개념이 사용되지 않음

- 대표적인 비지도 학습에는 k-평균군집, 계층적군집, 혼합분포군집을 비롯한 다양한 군집분석 과 주성분분석, 독립성분분석 등이 포함