Ⅰ. 금융 빅데이터와 데이터 분석 (Big data in finance & financial time series data analysis)

❸ 데이터마이닝 기술

대표적인 데이터 마이닝 기술

현재 상용 데이터 마이닝 소프트웨어에서 제공되는 주요한 알고리즘

- 연관 규칙(Association Rules)
- 순차 패턴(Sequential Patterns)
- 분류(Classification)
- 군집화(Clustering)
- 아웃라이어 판별(Outlier Discovery)

연관 규칙(Association Rules)

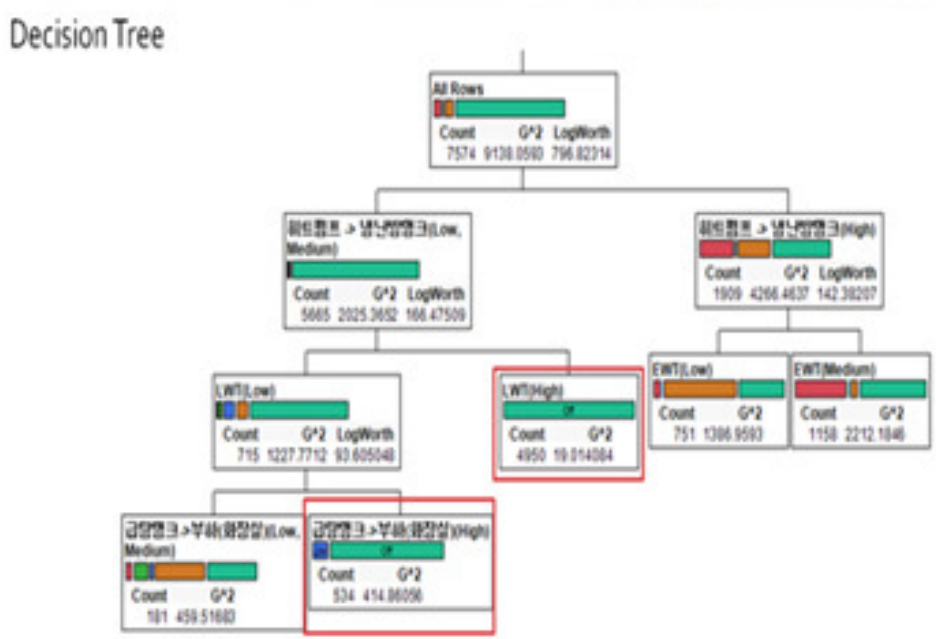
- 장바구니 분석
- 인터넷 쇼핑몰 및 오프라인 매장 등에서 고객이 한번에 구입하는 상품들을 분석하여 함께 판매되는 패턴이 강한 연관된 상품들을 찾는다.
- 예를 들어, [A.데이터마이닝 개론] 이라는 도서를 구입한 사람들은 [B.최신 마케팅 기술]이라는 교재를 함께 구입한다. 라는 패턴을 분석할 수 있고 이를 바탕으로 A도서를 구입한 고객에서 B도서의 구입을 추천할 수 있다.

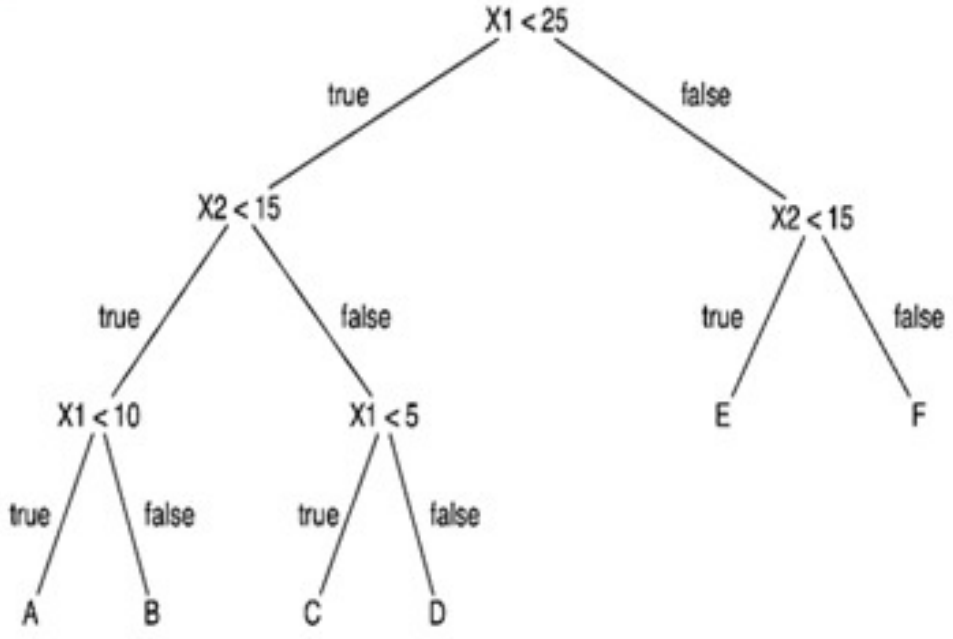
순차 패턴(Sequential Patterns)

- 연관규칙과 유사
- 연관규칙에 시간 정보를 추가하여 순차적인 구입 패턴을 분석하는 방법
- 예를 들어, 노트북을 구입한 사람들은 1달 정도 후에 노트북 받침대를 구입한다 라는 패턴을 찾을 수 있다. 이 규칙을 바탕으로 노트북을 구입한 고객들에게 노트북 받침대를 추천할 수 있다.

분류(Classification)

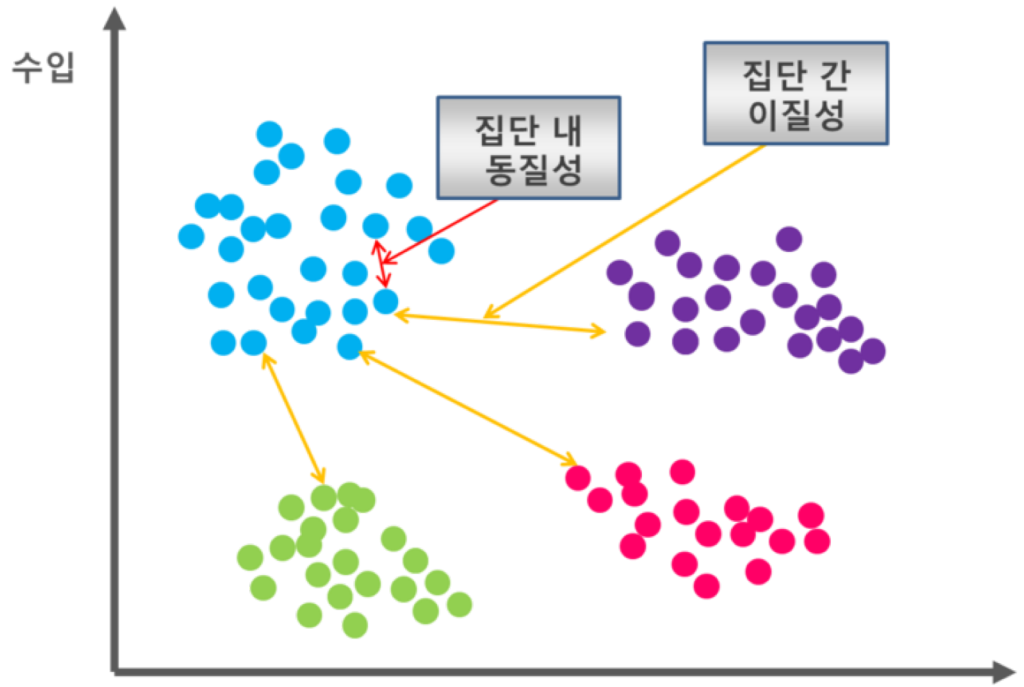
- 목표 필드의 값을 찾는 모델을 생성
- 과거의 데이터를 입력으로 하여 분류 모델을 생성하고 새로운 데이터에 대하여 분류 값을 예측





군집화(Clustering)

- 데이터를 여러가지 속성(변수)들을 고려하여 성질이 비슷한 몇 개의 집합으로 구분하는 분석 기법
- 분류 분석과는 달리 목표 변수를 설정하지 않는다.
- (따라서, 분류는 지도학습, 군집분석은 비지도학습이라고도 한다.)

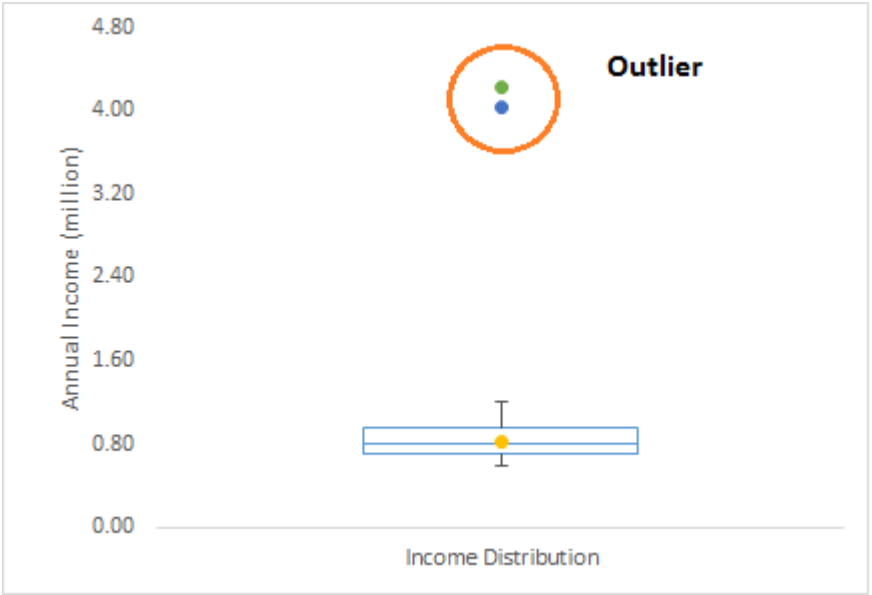


아웃라이어 판별 (Outlier discovery)

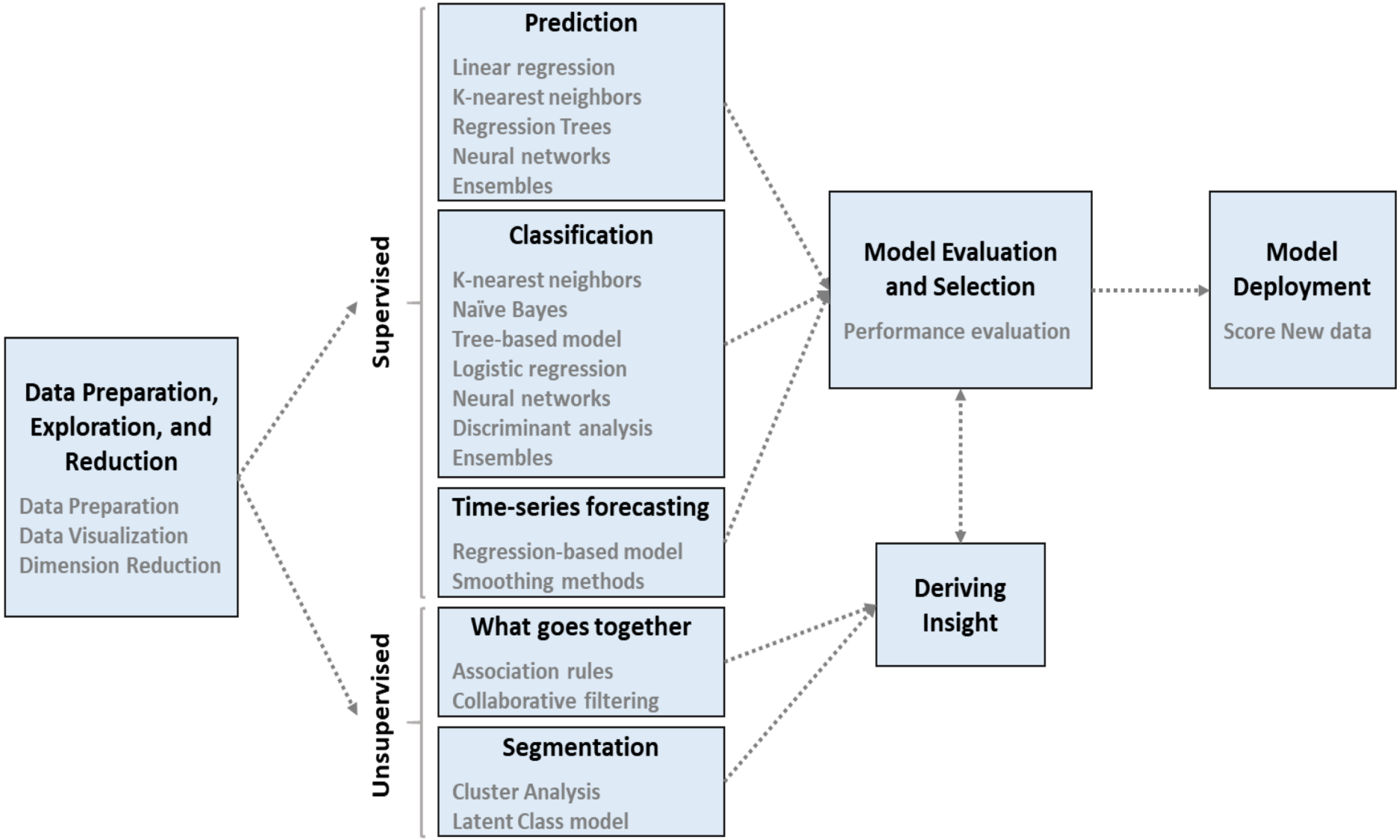
- 대부분의 데이터마이닝 기술은 데이터를 나타내는 패턴에 관심을 갖고 찾아내려 하는 반면에 아웃라이어 판별 기법은 이와 반대로 대부분의 데이터와 다른 소수 또는 일부를 찾아내는 기술
- 여러 유용한 곳에서 사용 가능
  - 전화카드를 훔쳐서 사용할 경우 자신의 카드를 사용하는 대다수의 선의의 고객들과 당연히 사용 패턴이 다름
  - 훔친 신용카드를 쓰는 사람들의 사용 패턴은 자기 카드를 사용하는 고객과 다를 수밖에 없을 것
  - 회사나 백화점 같은 곳에서 일반 고객의 동선과 도둑의 동선은 다를 것
  - 시스템에 침입한 크래커들이 사용한 명령어(command)들은 정상적인 사용자들과 다를 것

아웃라이어 판별 (Outlier discovery)

- 이러한 아웃라이어를 판별하는데 여러 가지 기술이 이미 통계학 분야에서 사용됐는데, 여기서 사용되는 알고리즘들은 주로 데이터를 일정한 통계적 분포(statistical distribution)로 가정해 모델을 설정하고 그 모델에 따라 아웃라이어를 판단하게 된다.
- 많은 경우에 사용자가 자신이 이용하려는 데이터의 분포를 알고 있지 않은 경우가 더 많다. 이를 위해 데이터베이스 분야에서 distance-based outlier discovery 알고리즘들이 개발되었다.



프로세스 관점의 데이터마이닝 기법 분류



Source : Shmueli et al. (2016). Data Mining for Business Analytics : Concepts, Techniques, and Application with XLMiner, 3rd Edition.

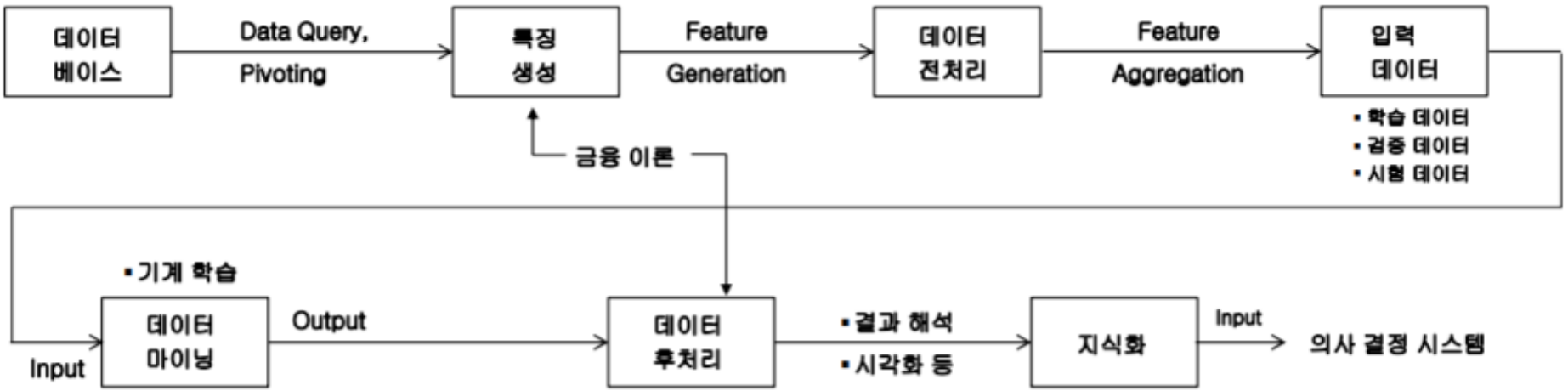
데이터마이닝 프로세스

- 지식 발견을 위해서는 데이터베이스로부터 분석 목적에 맞는 데이터를 선별하여 결과를 잘 설명할 수 있는 특징 (Feature) 집합을 생성함

  -> Feature Generation

- 특징 집합이 생성되면 데이터 표준화 등 전처리 과정을 거쳐 입력 데이터를 생성함

  -> Feature Aggregation



금융 데이터마이닝 프로세스

- 분석 목표가 설정되면 수집된 원시 데이터 (Trade & Market Data)로부터 목적에 적합한 특징 (Features) 혹은 속성 (Attribute)들을 생성함.

  -> **Feature Generation**

- 생성된 특징들을 모두 모아서 학습이 가능한 형태의 특징 집합을 생성함.

  -> **Feature Aggregation**

- 데이터 전처리 (Preprocessing) 과정을 거쳐 기계학습 알고리즘에 입력함.

금융 데이터마이닝 프로세스

상세 거래 데이터 :
- 호가, 잔량, 건수
- 거래 시각
- 체결 가격
- 체결 수량

요약 거래 데이터 :
- 요약 시간 (분봉, 일봉 등)
- 시가, 고가, 저가, 종가
- 거래량

시장 데이터 :
- 지수 (기초자산, 변동성 지수)
- 거래 주체별 수요, 공급량
- 파생 상품 관련

거래 데이터 자체 :
- 데이터 자체가 개별적으로 Features가 될 수 있음

미시 시장 Features :
- Depth, Liquidity, Spread ...
- Tick Time, Volume Time ...
- Order flow, Trade Intensity ...
- PIN, VPIN ...

기술적 분석 Features :
- MA, MACD, Bollinger Band
- Stochastic, RSI, CCI ...
- 변동성, 왜도, 첨도
- VWAP ... -
- 내재변동성, 민감도, 베이시스 ...

Feature 집합 생성
학습 가능한 형태로 가공
학습 데이터 세트 구성

데이터 정제
데이터 통합
데이터 변환
데이터 축소
데이터 이산화
전처리는 Trade & Market 데이터에서도 수행함.

기계학습
분류 (Classification)
군집 (Clustering)
회귀분석 (Regression)
연관분석 (Association)
추론 (Inference) 등.

과적합(Overfitting)

- **과적합 (Overfitting) 모델 :** 파라미터를 적당히 튜닝하지 않으면 모델의 정확도가 떨어진다.
- 예측 오류를 최소화하기 위해 예측 모델의 복잡도를 늘리면 예측의 경계선이 불필요하게 복잡해지는 문제를 초래할 수 있다.
- 모델의 복잡도를 적절하게 유지하는 방법 중의 하나로 정규화 단계에서 패널티 파라미터를 사용한다.
- 이 새로운 파라미터는 모델의 복잡도가 증가할 때 예측 오류를 인공적으로 키움으로서 복잡도 증가에 대한 불이익을 준다.
- 따라서 모델이 원래 파라미터를 최적화함에 있어 정확도와 복잡도를 모두 고려해야 한다.

최적합(Ideal Fit)

- **최적합 (Ideal Fit) 모델:** 파라미터를 제대로 튜닝하면 주요한 추세를 인식하는 일과 중요하지 않은 변동을 무시하는 일 사이에서 균형을 이루게 된다.

부적합(Underfit)

- **부적합 (Under fit) 모델 :** 민감도가 너무 떨어지며 숨겨진 패턴을 발견하지 못 함.
- 부적합된 모델은 중요한 추세를 놓치게 되고 이로 인해 현재 데이터와 미래 데이터 모두에서 예측 정확도가 떨어진다.