데이터마이닝(DataMining)

Chapter 1. 데이터마이닝의 개념

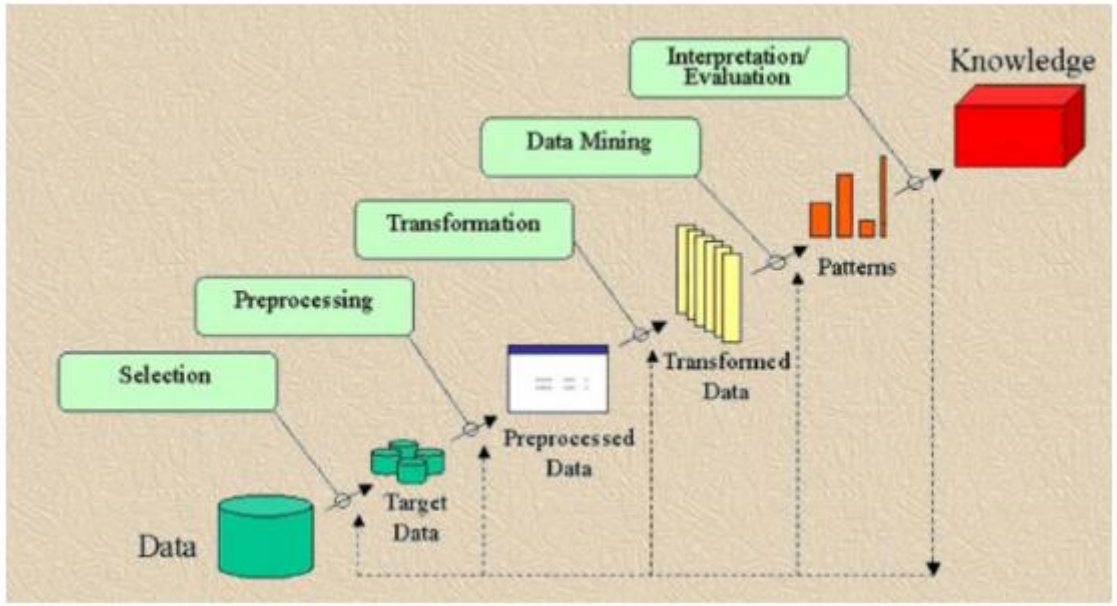
데이터마이닝이란?

- 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

데이터마이닝 적용 분야

- 데이터베이스 마케팅
  - 고객 자료를 분석하여 획득한 정보를 마케팅 전략 구축에 활용
  - 유통, 금융, 보건, 보험, 통신 등 다양한 부문에서 사용되며 데이터마이닝의 가장 성공적인 적용분야
  - 목표마케팅, 고객세분화, 고객성향변동분석, 장바구니분석 등
  - 추천시스템

- 신용평가
  - 특정인의 신용상태를 점수화하는 과정
  - 신용거래에서 대출여부와 대출한도를 결정
  - 불량채권과 대손을 추정하여 이를 최소함으로써 신용리스크를 관리하기 위해 사용
  - 금융기관에서 신용카드, 보험, 대출 등의 업무에 주로 사용

- 생물정보학
  - 마이크로어레이 기술은 세포의 수많은 유전자 발현값을 동시에 측정할 수 있게 함
  - 의료, 보건, 제약 등 바이오산업에 널리 활용
  - 인간 유전체 규명 계획으로부터 얻은 방대한 양의 유전자 정보로부터 가치 있는 정보를 추출해 질병의 진단과 치료법 또는 신약의 개발 등에 활용

- 텍스트마이닝
  - 전자우편이나 신문기사 등의 디지털화된 자료로부터 유용한 정보를 획득하는 과정
  - 입력 텍스트의 구조화, 구조화된 자료에서 패턴 검색, 결과에 대한 해석과 평가
  - 텍스트 분류, 텍스트 군집화, 개념/개체 추출, 세분화된 분류, 감성분석, 문서요약, 개체 관계 모형화 등에 적용
  - 인터넷 검색 엔진에서 관련 문서의 순위를 정하는 기술에 데이터 마이닝 기법 사용

- 사기방지
  - 신용대출 신청시 부정확한 정보 제공, 사기성 거래, 주민등록 번호 등을 이용한 신원절도, 보험 사기 등 여러 산업분야에서 문제로 화두
  - 사전에 사기행위를 탐지함으로써 회사와 고객이 사기행위에 노출될 위험을 줄이는 것을 목표
  - 예측 모형을 이용하여 자료로부터 사기행위 패턴을 분석하고 발견된 패턴을 데이터베이스 상에 규칙으로 저장하는 방법 등을 이용

❖ 통계와 데이터마이닝의 차이점 (1)

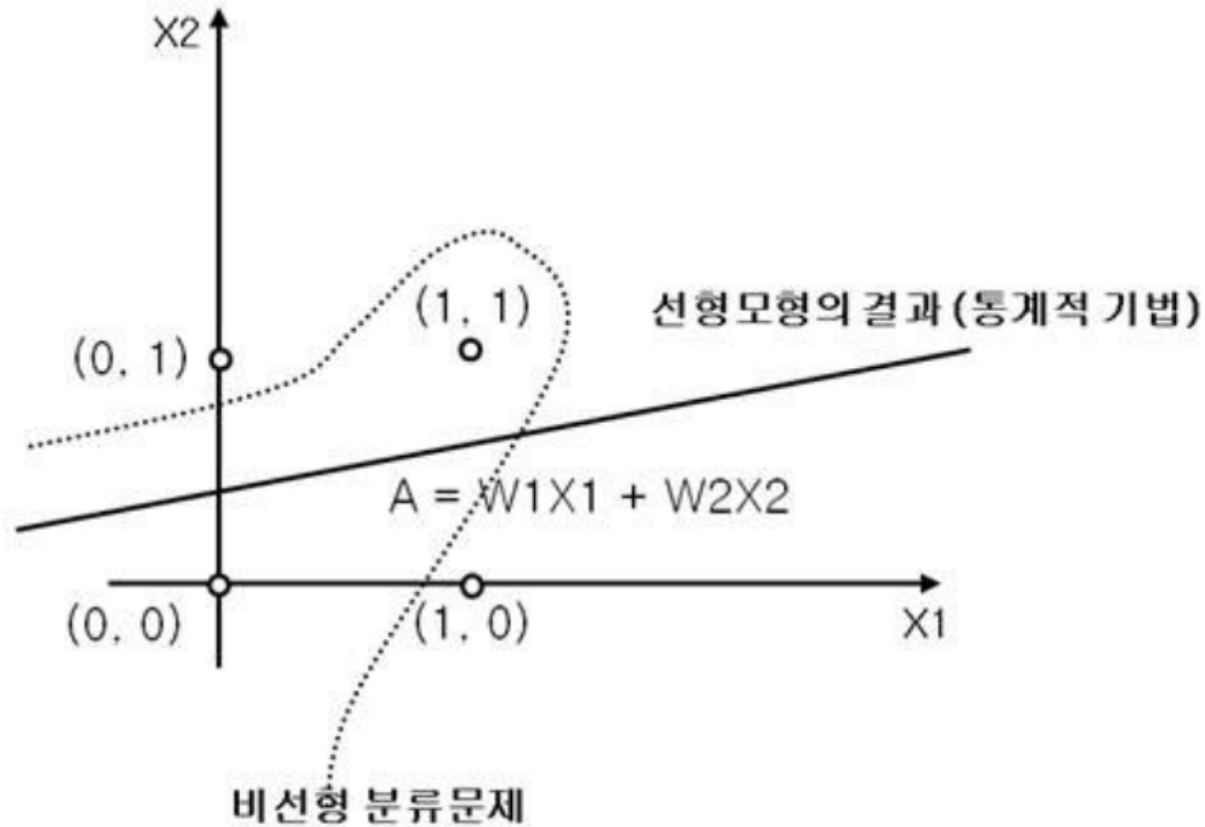
- ➢ 기존의 통계적 분석 도구나 OLAP은 세워진 모형이나 가설에 의거해 이를 검증하거나 요약 보고하는 데 초점을 맞추고 있는 반면, 데이터마이닝의 목적은 궁극적으로 예측에 초점을 둠
- ➢ 데이터마이닝에 사용되는 인공 지능 기법은 그 어떠한 기법보다 모형의 예측 성과를 높이는데 가장 우수한 기법임

전통적인 통계	데이터마이닝
• 현실에 적용하기 부적합한 가정 (Assumption) • 예) 모집단의 정규분포, 선형성, 등분산성 등 • 제안된 가설에 대한 검증이 주 목적 • 알고리즘이 선형성에 기반을 두고 있음	• 현실적인 noisy한 데이터에 대한 가정이 없음 • 알고리즘이 비선형성에 기반을 두고 있음 • 미래를 예측하는 것이 주 목적 • 모형에 대해 Robust한 결과를 제공함 • 예측 성과가 통계기법보다 우수한 것으로 많은 실증연구에서 검증되었음

❖ 통계와 데이터마이닝의 차이점 (2)

현실의 데이터는 대부분 비선형이며
따라서 비선형 문제를 해결하기 위해서는
비선형 모형이 필요함

“인공지능 알고리즘을 이용한 데이터 마이닝”



❖ 통계와 데이터마이닝의 차이점 (3)

- 데이터마이닝은 통계학과 기계학습(machine learning: 인공지능으로도 알려짐)으로 알려진 두 학문분야의 합류점에서 존재
- 데이터를 탐색하고 모델을 구축하는 다양한 기법들은 통계학분야에서 오랫동안 존재해옴
- 예를 들어 여기에는 선형 회귀분석, 로지스틱 회귀분석, 판별분석, 주성분 분석 등이 포함
- 그러나 충분한 데이터와 계산능력을 가진 데이터마이닝의 응용분야가 적에서는 이러한 고전적인 통계학의 핵심원리(계산이 어렵고 데이터가 희소하다는 것) 용되지 않는 것이 특징

Daryl Pregibon은 데이터마이닝을 규모와 속도의 통계학으로 묘사함(Pregibon, 1999)

- 통계학 분야와 구별되는 특징
  - 실험계획이나 샘플링에서는 사전 계획에 따라 자료가 수집되지만, 데이터 마이닝에서는 대용량의 관측 가능한 자료를 다룸. 변수사이의 관계를 왜곡할 가능성 존재
  - 분석하는 자료가 모집단인 경우가 많음. 경험적 방법론들을 매우 중시
  - 미래에 대한 예측을 중시

- 데이터 베이스에서 지식탐색
  - 데이터마이닝의 결과를 이용하여 유용하며 이해할 수 있는 정보로 변환하는 것
  - 자료의 전처리, 데이터마이닝, 결과의 검증 및 해석
  - 전처리는 노이즈나 결측치를 제거하거나 자료를 변화하여 분석에 사용될 수 있도록 자료를 정리하는 것
  - 검증 및 해석은 데이터마이닝 결과에서 찾은 패턴이 모집단의 다른 부분으로 일반화가 잘 되는지 여러 가지 방법의 예측력을 비교하고 그 결과에 대한 해석

- 기계학습
  - 인공지능의 한 분야로서 컴퓨터로 하여금 센서나 데이터베이스로부터의 자료를 이용하여 판단하도록 하는 알고리즘을 개발하는 학문 분야
  - 자료에 기반하여 복잡한 패턴을 인식하고 의사결정을 내릴 수 있게 자동적으로 학습하는데에 중점
  - 컴퓨터비전, 웹 검색엔진, 음성 및 문자인식, 게임, 로봇운동 등

- 패턴인식
  - 원자료를 이용하여 사전지식이나 패턴에서 추출된 통계적 정보에 기반하여 자료 또는 패턴을 분류하는 것이 목적
  - 분류될 패턴은 흔히 다차원 공간에서 점으로 정의되는 관측치의 그룹

데이터마이닝 적용 사례

- 유통
  - 미국의 한 할인점에서는 매장내의 상품진열과 고객들의 구매패턴의 연관성을 발견하기 위해 연관성 분석을 적용
  - 기저귀와 맥주가 강한 연관성을 보였고 매장에서 기저귀와 맥주를 가까이 배치함으로써 매출이 증가
- 금융
  - 신용카드회사는 사기행위를 적발하고 이를 예방하기 위해 정상적인 거래와 사기행위 거래 자료를 의사결정나무와 신경망 등을 이용하여 분석하여 사기방지 시스템을 구축
  - 사기방지 시스템에서는 승인을 요청하는 거래의 패턴이 카드 소지자의 기존 패턴과 다른 경우 부정사용으로 의심하여 거래에 대한 승인을 보류하고 콜센터에서 카드소지자에게 전화로 거래여부를 확인
  - 고객의 자산 보호 및 회사의 손해액 감소

- 의료
  - 미국의 한 대학병원에서는 종양 진단의 정확성을 높이기 위해 종양의 악성 또는 양성 여부를 판별 및 분류분석을 시행
  - 과거의 환자들에 대한 종양검사의 결과를 근거로 종양의 악성 또는 양성 여부를 예측하는 분류모형을 만든 후, 새로운 환자에게서 얻은 측정값을 사용하여 악성 또는 양성 여부를 판별
  - 얻은 결과를 진단 시 참고자료로 활용
  - 최근에는 유전자 칩을 이용한 암진단 분야에 대한 관심 고조

- 제조
  - 반도체회사에서는 정상인 반도체를 그 특성에 기반하여 몇 개의 군집으로 나눈 후, 새로운 제품이 정상제품들의 군집 범위의 밖에 이는 경우 불량으로 규정
  - 연관성 분석과 군집분석 사용
  - 반도체 제품의 불량율 감소
- 통신
  - 통신사에서는 매년 23%의 고객이 이탈하는데 새로운 고객 한 명을 유치하는데 비용 발생
  - 고객성향변동관리와 군집분석을 이용한 결과 현재 고객의 40%거 이탈 가능성이 높은 것으로 나타남
  - 이익분석을 통해 이탈가능성이 높은 고객들 중 회사의 이탈방지 노력이 효과적일 것으로 예상되는 고객을 선별하여 통화서비스 등의 목표마케팅으로 이탈고객이 19.7%로 감소

- 정보산업
  - 어느 회사에서는 개별 고객의 과거 구매제품과 관련있는 상품을 발굴하고 이를 고객에게 개별적으로 소개
  - 인터넷 DVD 대여 회사는 개별 고객의 취향을 고려한 영화를 추천하는 시스침을 운영
  - 검색 사이트는 각 질의(query)에 대응되는 관련성이 높은 문서의 탐색을 위해 데이터 마이닝 기법을 사용

데이터마이닝의 솔루션

- SAS Enterprise Miner
- SPSS Clementine
- IBM Intelligent Miner
- Oracle Darwin
- Salford CART&MARS
- R