데이터마이닝의기본개념

Jinseog Kim (jinseog.kim@gmail.com)

2017년 3월

데이터 마이닝이란

- 정의: 대용량의 데이터 내에 존재하는 **관계**, **패턴**, **규칙** 등을 탐색하고 모형화함으로써 유용한 지식을 추출하는 과정
- 배경
  - 지식정보화 사회에서는 새로운 지식의 습득이 경쟁력의 원천 (예: 유전자 정보, 고객 정보 등)
  - 거대한 데이터의 분석을 통하여 새로운 지식 발견 가능
  - 자료의 효율적 저장을 위한 기술 (데이터 베이스, 압축, 통신)의 발달에 의한 방대한 양의 데이터 집적
  - 컴퓨터 성능의 향상으로 인하여 거대한 데이터의 실시간 분석 가능

데이터 마이닝의 활용분야 1

- 데이터베이스 마케팅(Database Marketing)
  - 데이터를 분석하여 획득한 정보를 이용하여 마케팅 전략 구축
  - 예:
    - 목표마케팅(Target Marketing)
    - 고객 세분화(Segmentation)
    - 이탈고객분석(Churn Analysis)
    - 장바구니 분석(Market Basket Analysis)
    - 상품추천 등...
- 신용평가 (Credit Scoring)
  - 특정인의 신용상태를 점수화하는 과정
  - 신용거래 대출한도를 결정하는 것이 주요 목표
  - 이를 통하여 불량채권과 대손을 추정하여 최소화함
  - 예: 신용카드, 주택할부금융, 소비자 / 상업 대출
- 생물정보학 (Bioinformatics)
  - 지놈(Genome)프로젝트로부터 얻은 방대한 양의 유전자 정보로부터 가치 있는 정보의 추출
  - 응용분야: 신약개발, 조기진단, 유전자 치료

데이터 마이닝의 활용분야 2

- 텍스트 마이닝 (Text Mining)
  - 디지털화된 자료 (예: 전자우편, 신문기사 등)로 부터 유용한 정보를 획득
  - 응용분야: 자동응답시스템, 소셜미디어 분석, 상품평 분석, 전자도서관, Web surfing
- 부정행위 적발 (Fraud Detection)
  - 고도의 사기행위를 발견할 수 있는 패턴을 자료로부터 획득
  - 응용분야 : 신용카드 거래사기 탐지, 부정수표 적발, 부당/과다 보험료 청구 탐지

데이터마이닝의 특징

- 대용량의 관측 가능한 (주로 비계획적으로 수집된) 자료를 다룸
- 컴퓨터 중심의 기법
- 경험적 방법이 중시됨
- 일반화 (generalization) 또는 예측이 중요 : 현재의 자료보다 미래의 자료를 잘 설명할 수 있는 모형을 추구
- 통계학과 컴퓨터공학(특히 인공지능)에서 함께 방법론을 개발하고 이를 경영, 경제, 정보기술(IT)분야에서 사용

데이터마이닝 관련 분야

- KDD (Knowledge Discovery in Database)
  - 데이터베이스안에서의 지식발견
  - 데이터마이닝과 가장 유사
  - KDD는 지식을 추출하는 전 과정 (계획, 자료 획득, 분석, 해석 등)을 의미하고 데이터마이닝은 KDD의 한 과정(자료의 분석)임
  - 데이터 웨어하우징 (data warehousing), OLAP (On-Line Analytical Process-ing) 등도 KDD의 한 과정
- 기계학습 (Machine Learning)
  - 인공지능 (Artificial intelligence)의 한 분야
  - 입력되는 자료를 바탕으로 기계(컴퓨터)가 판단을 할 수 있는 방법에 대한 연구
- 패턴인식 (Pattern Recognition)
  - 거대한 자료로부터 일정한 패턴을 찾아가는 과정
  - 이미지 분류와 깊은 관련이 있음
  - 통계학의 판별(분류) 분석, 군집분석 등이 사용됨
- 통계학
  - 많은 데이터마이닝 기법들은 통계학 관점에서 비선형 함수추정 문제임
  - 예: 신경망 모형 대 로지스틱 or 사영추적회귀(Projection Pursuit Regression)

데이터마이닝 적용사례 1

- 소매/유통업
  - 미국의 할인점 Wall Mart에서 매장내의 상품들과 고객들의 구매패턴의 연관성을 발견하기 위하여 연관성 분석 알고리즘을 사용
  - 기저귀와 맥주가 강한 연관성을 나타냄
  - 기저귀와 맥주를 가까이 배치하여 매출이 증가
- 신용카드회사
  - 국내의 한 신용카드회사가 부정행위를 적발하고 이를 예방하기 위한 모형의 구축
  - 기존의 카드 소지자의 구매패턴을 분석하여 현재의 구매패턴이 카드 소지자의 구매패턴과 틀린 경우 부정사용으로 의심
  - 의사결정나무와 신경망 모형, 딥러닝 기법등이 사용됨
  - 카드의 부정사용 방지를 통하여 고객의 자산 보호 및 회사의 손해액 감소

데이터마이닝 적용사례 2

- 의료분야
  - 종양의 악성/양성 판단에 의한 암 진단의 정확성을 높이기 위한 판별 및 분류분석 시행
  - 과거의 환자들에 대해서 종양검사의 결과를 근거로(즉, 종양의 크기, 모양, 색깔 등) 종양의 악성/양성 여부를 구별하는 분류모형을 만든 후, 새로운 환자에서 얻은 입력변수를 이용하여 암을 진단
  - 지도학습방법 (신경망, 로지스틱 회귀모형, 의사결정나무 등)이 사용
- 제조업
  - 반도체회사에서 불량품 자동검색장치 개발
  - 연관성 분석과 군집분석 알고리즘을 사용
  - 정상인 반도체를 그 특성에 기반하여 몇 개의 군집으로 나눈 후, 새로운 제품이 정상제품의 군집의 범위밖에 있는 경우 불량으로 규정
  - 불량품 감소로 인한 이익의 증대

데이터마이닝 적용사례 3

- 통신회사
  - 미국의 한 장거리 통신 회사의 23%의 고객이 매년 이탈
  - 새로운 고객 한명을 유치하는데 필요한 비용이 $350
  - 이탈고객관리(churn management)와 군집분석(clustering)을 이용하여 이탈의 원인을 파악 현재 고객의 40%가 이탈 가능성이 높음
  - 이익분석(profit analysis)를 통하여 이탈가능성이 높은 고객을 상대로 한 마케팅이 효과적임이 입증
  - 무료 통화서비스 등의 목표마케팅(target marketing)으로 이탈고객 감소와 이를 통한 이익의 증가

데이터마이닝의 기본 방법론

1. 회귀분석(regression analysis): 선형회귀모형, 변수선택, 로지스틱 회귀
2. 모형평가(model assessment): 모형들의 예측성능을 평가하는 방법
3. 의사결정나무(decision trees)
4. 신경망(neural networks)
5. 기타 지도학습기법: 단순 베이즈 분류(naive Bayes classifier), k-근방 분류, SVM
6. 차원축소기법: 주성분분석(principal component analysis), 인자분석(factor analysis), 다차원 척도법(multidimensional scaling)
7. 연관규칙분석(association rule analysis)
8. 군집분석(cluster analysis): k-평균군집법(k-means clustering), 계층적 군집법(hierarchical clustering)