1장

데이터마이닝(Data Mining)

1.1 데이터마이닝이란?

▸ 대량의 데이터를 분석하여 중요한 정보를 얻어 현실에 응용하는 기법을 의미
예) 웹로그(web log) 데이터 분석: 젊은 사람들이 제일 많이 이용하는 웹사이트는 무엇인가? 이 웹사이트 중에서 어떠한 웹페이지가 가장 많은 관심의 대상인지 파악하여 이들을 대상으로 하는 광고 전략을 세움

핸드폰 통화기록 데이터
은행거래 데이터: 은행 자동화기기(ATM)를 이용할 때마다 생성
개인별 구입상품 목록 데이터: 슈퍼마켓에서 상품을 구입할 때 상품의 bar code를 판독하여 생성

▸ 데이터의 양이 커서 기존의 분석방법을 사용할 수 없거나 별로 의미가 없는 경우가 많음
▸ 데이터마이닝: 전통적인 데이터 분석기법에 대량의 데이터 처리기법을 가미하여 발전시킨 데이터 분석기법
▸ 마이닝(mining): 광산(mine)에서 광물(valuable minerals) 채광(extraction)한다는 뜻
▸ 데이터마이닝: 대량의 데이터라는 거대 광산에서 유용한 보물(정보)를 찾는다는 뜻
⇒ 지식발견(knowledge discovery)
발견된 정보 ⇒ 의사결정에 이용, 예측, 과거에 분석했던 데이터를 다시 조명해보는 기회를 제공

▸ 데이터마이닝은 여러 학문 분야와 연관이 되어 있음 (그림 1-1 참조)
(1) 데이터마이닝: 데이터 분석 & 예측모형
(2) 통계학: 표본론(Sampling), 추정(Estimation), 가설검정(Hypothesis test), 다변량분석(Multivariate analysis), Bootstrap and CART(Classification analysis and regression tree) 등이 사용됨
(3) 컴퓨터과학: 다량의 데이터를 효율적으로 처리하기 위해 컴퓨터과학의 데이터베이스 시스템(database system), 분산 컴퓨팅(distributed computing) 이론, 병렬 컴퓨팅(parallel computing) 이론, 기계학습(machine learning) 이론이 사용됨
(4) 수학: 최적화 이론(optimization theory), 신호처리 이론(signal processing theory), 정보 이론(information theory), 정보 검색(information retrieval), 경영정보 시스템(management information system)이 사용됨
(5) 그 외 생물학의 유전 이론(genetics) 이용
▷ 통계학자나 컴퓨터 학자 또는 경영과학자들이 서로 자기 분야라고 주장함
⇒ 모두 맞는 말
⇒ 여러 학문분야의 사람들이 협력하여 데이터마이닝이라는 새로운 분야를 만들었음

1.2 데이터마이닝 응용분야

가) 기업의 경영

▷ 고객관계 경영(customer relation management: CRM)
대형 슈퍼마켓, 백화점에서 자체 신용카드 발급 → 신용카드 이용현황 분석
→ 고객의 상품 구매성향 정보 얻음
→ 이런 정보를 분석하여 특정한 부류의 고객만을 대상으로 하는 표적마케팅(target marketing) 전략을 세울 수 있음

▷ 대형 슈퍼마켓에서 쇼핑 → 계산대에서 바코드 scan
→ 고객이 구매한 상품 목록과 총 구매액수 계산 →데이터마이닝 기법 적용하여 분석
→ 고객들이 어떠한 상품의 구매에 관심이 많은지 또는 구매할 때 서로 연관이 있는 상품이 어떤 것인지 알 수 있음: 장바구니 분석(market basket analysis)

나) 유전자 분석

▷ 특정한 병의 원인이 되는 유전자를 찾아 질병치료에 이용

▷ 유전자 데이터는 다차원이어서 이를 분석하기 위해 데이터마이닝의 기법이 이용됨

다) 지구과학

▷ 과학위성들이 지구로 보내는 데이터: 공간적(spatial)이고 시간적(temporal)인 주기를 가짐. 데이터양이 워낙 방대함

▷ 데이터마이닝 분석기법이 이용됨

라) 정보과학

▷ 대학의 도서관이나 공공도서관에 가면 사용자가 찾고자하는 정보의 주요 단어를 시스템에 입력하면 문서나 책 검색: 키워드 정보 검색(keyword information retrieval)

▷ 하지만 너무 많은 문서 검색 - 불필요한 문서도 검색

▷ 좀 더 효율적인 방법으로 본인이 원하는 정보를 검색하는 방법이 연구되고 있음
: 텍스트 마이닝(text mining)이라 함

1.3 데이터마이닝 과정 (그림 1-2 참조)

▶ 입력데이터 형태
- 텍스트 파일, 스프레드시트 파일, 여러 개의 데이터베이스에 흩어져 있는 데이터

▶ 데이터 탐색 및 전처리(preprocessing): 데이터를 탐색하고 데이터마이닝을 위해 적당한 형태로 변환 (Chap 2)
- 잡음(noise) 데이터, 중복 데이터 제거, 데이터 통합
- 정규화(normalization), 이산형화(discretization), 변환(transformation), 일부 데이터 추출, 차원 축소 변환

▶ 데이터마이닝 모형화(modeling): 주어진 문제에 적합한 분석방법 결정 (Chap 5, 6, 7)
- 연관분석(association analysis), 분류분석(classification), 군집분석(clustering)을 이용하여 모형 수립

▶ 분석결과 후처리(postprocessing): 데이터마이닝 결과를 의사결정에 효율적으로 사용하기 위해 시각화(visualization)하거나 의미 있는 분석결과만 추출, 요약, 설명

▶ 유용한 정보