제 1장
데이터마이닝의 주요 개념

INDEX

1.1 데이터마이닝이란 무엇인가?

❖ Data mining : database, data warehouse, data mart 등 자료저장소에 저장되어 있는 방대한 양의 데이터로부터 의사결정에 도움이 되는 유용한 정보를 발견하는 일련의 작업들의 집합.

금(金)광산 -> 채굴(mining) -> 금(Gold)

Data warehouse -> ( ) -> information(정보)

1.1.1 정보기술의 발전과 데이터 마이닝

정보기술(information technology)분야의 기술 발전

-> 대용량의 데이터를 축적하고 분석하는 것이 가능

-> 데이터 관리, 분석을 위해 다양한 형태의 데이터베이스 시스템 출현

-> 우수한 성능의 데이터마이닝 소프트웨어들의 사용 가능

1.1.2 빅데이터(Big Data) 분석

- 빅데이터 : 데이터의 생성 양, 주기, 형식 등이 기존 데이터에 비해 너무 크기 때문에, 종래의 방법으로는 수집, 저장, 검색, 분석이 어려운 방대한 데이터
- 기업의 빅데이터 활용 : 고객의 행동을 미리 예측하고 대처방안을 마련해 기업 경쟁력을 강화, 생산성 향상과 비즈니스 혁신

1.1.3 고객관계관리(Customer R_____ Management)

- 기업경영에서는 고객의 요구가 무엇인지 지속적으로 파악하는 것이 중요 -> 고객 정보를 지속적으로 축적하고 분석해야 함
- 고객관계관리(CRM) : 기업과 고객간의 상호교류를 관리하는 프로세스 -> 고객정보를 분석하는 도구로서 데이터마이닝의 사용이 요구

1.1.4 데이터마이닝 관련 분야

- 데이터베이스로부터의 지식발견(KDD ; Knowledge Discovery in Database)

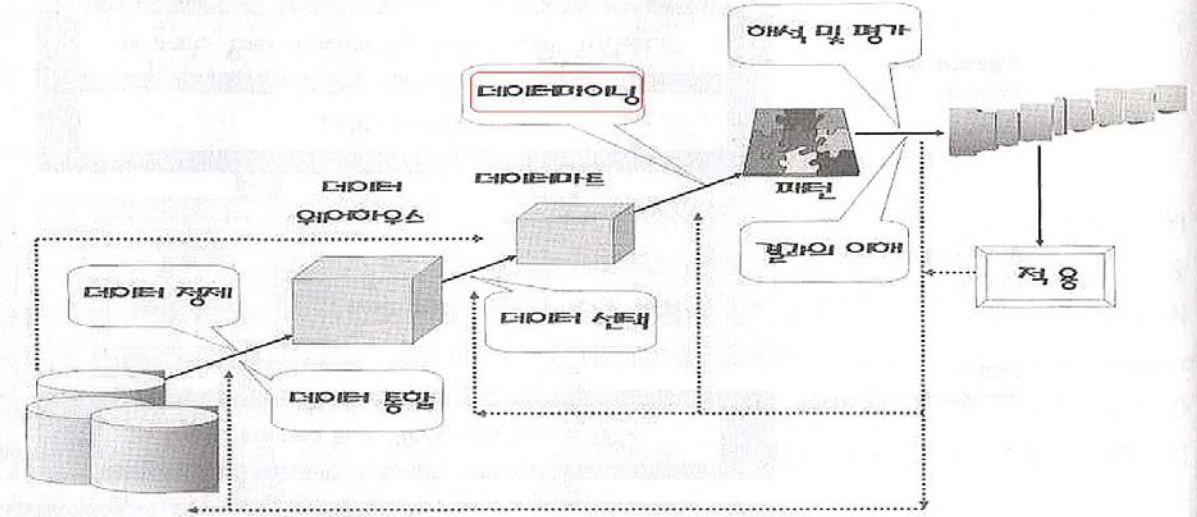

그림 1.2 데이터베이스로부터의 지식발견(KDD) 과정

- 기계학습 (Machin Learning) – 자동적인 학습 기법 설계, 구현
- 패턴인식 (Pattern Recognition) – 문자인식, 이미지분류
- 통계학 (Statistics) – 다변량 (판별분석, 주성분분석, 군집분석), 회귀분석 등
- 뉴로컴퓨팅 (Neurocomputing) - 신경망분석

1.1.5 데이터마이닝의 활용분야와 특징

❖ 데이터마이닝의 활용분야

- 데이터베이스 마케팅 (Database Marketing)

  : 목표마케팅, 고객세분화, 고객성향변동분석, 장바구니 분석 등

- 신용평가 (Credit Scoring) – 불량채권과 대손 추정하여 최소화
- 품질관리 (Quality Control) – 불량품을 찾아 원인을 찾아 예방

Define -> Measure -> Analyze -> Improve -> Control

- 부정행위의 적발 (Fraud Detection) – 사기행위 패턴 발견하여 예방
- 이미지분석 (Image Analysis) – 디지털 데이터로 부터 패턴 추출

❖ 데이터마이닝의 특징

- 대용량의 관측 가능한 자료를 다룬다.
- 관측자료는 시간의 흐름에 따라 비계획적으로 축적되며,

  자료분석을 염두에 두고 수집되지는 않는다.
- 컴퓨터 중심적 기법이다.
- 수리적으로 밝혀지지 않는 경험적 방법에 근거하고 있다.
- 일반화에 초점을 두고 있다.
- 경쟁력 확보를 위한 의사결정을 지원하기 위해 활용된다.

1.1.6 데이터마이닝 적용 사례

- 소매업 적용사례 - 고객의 구매패턴 -> (연관성분석 장바구니분석)
- 신용카드회사 적용사례 – 부정행위적발, 예방

  -> (의사결정나무분석, 신경망분석)
- 의료분야 적용사례 – 암진단 -> (판별 및 분류분석)
- 제조업 적용사례 - 불량품의 자동발견 -> (연관성규칙분석, 군집분석)
- 통신회사 적용사례 – 고객전화사용패턴 -> (군집분석)
- 스포츠경영 적용사례 – (소비자에 대한 마케팅전략)

1.2 데이터마이닝 프로젝트의 수행 프로세스

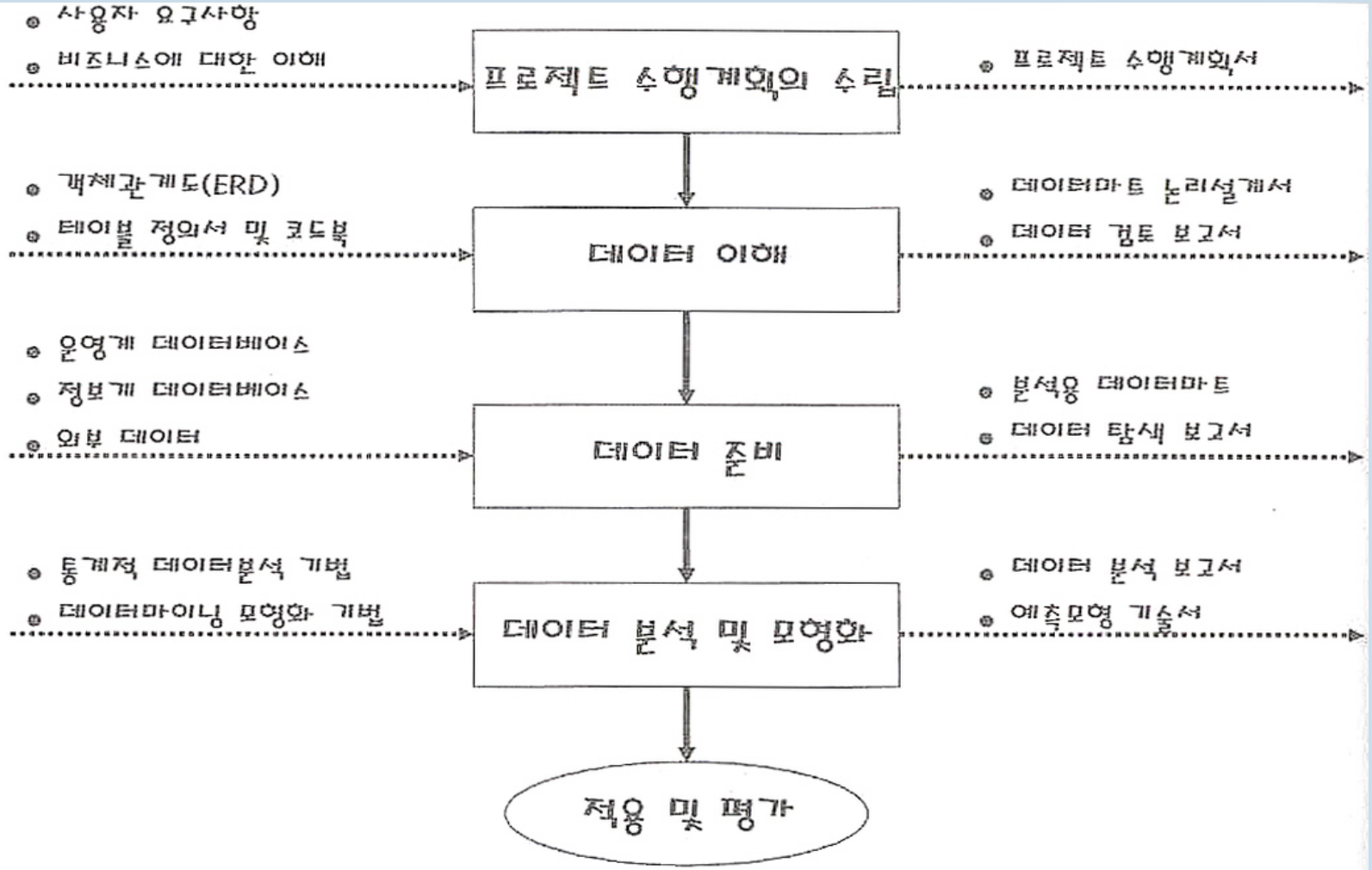

그림 1.5 데이터마이닝 프로젝트의 수행 프로세스

1.2.1 프로젝트 수행계획의 수립

❖ 성공적인 데이터마이닝 프로젝트 수행을 위한 요소

- 해당 비즈니스에 대한 충분한 이해
- 필요한 데이터를 관리하고 추출할 수 있는 정보기술
- 적절한 데이터 처리와 분석을 수행할 수 있는 통계적 분석능력

❖ 검토사항

- 프로젝트의 범위와 산출물 정의
- 비즈니스에 대한 이해 및 공유
- 사용자 요구사항과 필요사항
- 참여 인력 및 역할에 대한 정의
- 세부일정 정의 및 수행계획서 작성

1.2.2 데이터에 대한 이해

- 데이터마이닝의 성공여부 -> 사용 가능한 데이터의 양과 질에 의존
- 사용 가능한 데이터를 검토하고 특징을 이해하는 것이 첫단계
- 데이터 마이닝 작업을 수행하기에 앞서 데이터 원천들을 파악,
- 어떤 데이터를 수집하고 이용할 수 있는지 검토

❖ 검토사항

- 사용 가능한 내부 및 외부 데이터들의 원천
- 데이터 원천들에 대한 위치와 구조
- 데이터 테이블들의 필드와 코드
- 데이터의 신뢰성, 정확성, 유용성
- 분석용 데이터마트를 구성하기 위한 논리설계서 작성

❖ 데이터의 유형

✓ 서술적 데이터 (Descriptive Data)

- 개인이나 가구의 특성 묘사, 보통 요약 데이터의 형태.
- 고객의 기본 정보라서 자주 변하지 않음.
- 안정적이라 예측 모형을 구축하는데 유용.
- 응답자가 쉽게 응답하지 않거나 거짓 정보 제공.

✓ 행동특성 데이터 (Behavioral Data)

- 기업이 고객과 상호교류함으로써 발생하는 데이터.
- 고객의 행동이나 행위를 측정한 것이라 예측모형에 유용.
- 시간에 따라 빠르게 변화하며 쉽게 갱신가능.

✓ 태도특성 데이터 (Attitudinal Data)

- 고객의 태도 또는 심리적 특성을 측정한 데이터.
- 정확한 데이터 수집이 어려움.

❖ 데이터의 원천

✓ 운영계 데이터베이스 (Operational Database)

- 기업의 운영과 관련된 업무처리를 위해 구축된 것.
- 최근의 데이터를 저장하며 대량의 데이터 저장에 용이

✓ 정보계 데이터베이스 (Informational Database)

- 정보분석을 위해 구축된 것.
- 수집된 데이터를 요약, 가공하여 저장.

✓ 데이터 웨어하우스 (Data Warehouse)

- 조직 전체를 통해 데이터에 대한 통합된 관점 제공.
- 좀 더 적절하고 유용한 정보를 만들 수 있도록 데이터 요약

✓ 데이터 마트 (Data Mart)

- 하나의 데이터마이닝 주제 또는 고객분석을 위해 통합된 데이터로 구성된 보조적인 데이터 저장소.
- 특정한 목적의 사용자 그룹을 위해 특정 주제영역의 데이터로 구성.

✓ 메타 데이터 (Meta Data)

-데이터에 대한 데이터

- 데이터 관리를 위해 매우 상세하고 이해하기 쉽게 작성되어야 함.

1.2.3 데이터 준비 : 분석용 데이터마트 만들기

❖ 데이터 사전처리(Pre-processing of data)

서로 다른 목적을 가지고 수집된 상이한 유형의 데이터를 대상으로 하므로, 하나의 데이터 마트로 통합하기 위해 사전작업이 필요.

- 재배열(Rearrangement)

고객	구매일	상품	비고
3135	970304	A01	
3135	980715	B01	
3135	991113	C01	
2784	930508	C02	
2784	980106	B01	
8321	910305	A02	

고객	P_A	P_B	P_C
3135	1	1	1
2784	0	1	1
8321	1	0	0

- 요약변수(Summary Variable)

고객	구매일	상품	금액
3135	970304	A01	160
3135	980715	B01	42
3135	991113	C01	212
2784	930508	C02	250
2784	980106	B01	122
8321	910305	A02	786

고객	총금액	평균금액	건수
3135	414	138	3
2784	272	136	2
8321	786	786	1

최근 6개월 구매건수
최근 12개월 구매건수
최근 6개월 구매금액
최근 12개월 구매금액
…

- 파생변수 : 기존의 변수들을 이용하여 만들어진 새로운 변수

Ex) 식료품과 의류제품 구입금액의 합계, 총구매금액 대비 전자제품 구매금액 등과 같이 비율, 합계, 평균을 계산하는 작업이 포함

- 그룹화 : 변수의 차원을 축소하여 보다 의미 있는 해석이 가능

Ex) 범주형 변수를 다시 묶어서 범주의 수를 줄이는 재그룹화,

연속형 변수를 구간화하는 것 등, 신중년 패턴은 몇세부터 ?

❖ 데이터에 대한 탐색 및 보완

- 오류값 : 변수가 가질 수 없는 값, 잘못된 코드값 등 -> 오류를 파악하여 적절한 값으로 변경 or 데이터 표준화(정규화)를 통해 오류 수정

- 결측값 : 원인과 기록방법을 조사하여 자료를 정정하고 기록방법 변경 (단일값 대체, 클래스 대체, 다변량적 대체 등)

- 이상치 : 분석의 목적에 적합하지 않은 특이한 개체를 예측모형 구축에서 제외하거나 이상치를 적절한 값으로 대체

1.2.4 데이터 분석 및 모형화

- 만들어진 데이터마트를 이용하여 데이터에 대한 분석 및 예측모형의 구축 수행

- 다양한 평가도구들을 이용하여 예측모형 평가, 최종 예측모형 결정

1.2.5 적용 및 평가

- 다른 프로세스와 유기적으로 연관되도록 적용.

- 지속적 관리, 평가 필요

1.3 데이터마이닝 예측기법

1.3.1 지도예측(Supervised Prediction)

- (입력변수, 목표변수가) 존재
- 입력변수로부터 목표값을 예측하는(인과관계) 모형 개발이 목적
- 목표변수의 형태에 따라 두 가지로 나뉨.

  ① 범주형 목표변수 ② 연속형 목표변수
- (판별분석, 회귀분석, 의사결정나무분석, 신경망분석, 시계열분석)등

1.3.2 자율예측(Unsupervised Prediction)– 비지도예측

- 목표변수가 명확히 규정되지 않음 (입력변수만 존재).
- 데이터에 존재하는 여러 형태의 특징을 찾는 것이 목적.
- (주성분분석, 요인분석 군집분석, 연관성분석) 등

1.4 Enterprise Miner의 소개

1.4.1 Enterprise Miner(e-miner) 의 장점

- 분석흐름도(PFD ; process flow diagram) 와 노드(node)를 이용해 데이터 마이닝의 전 과정을 GUI환경에서 쉽게 수행가능.
- 다양한 통계분석, 신경망, 의사결정나무 등 강력한 분석기법 제공.
- 다양한 모형 평가 기능을 이용해 유용성과 적합성 파악이 쉬움.
- 언제든 재실행이 가능.
- 데이터 관리 기능이 우수하고 파생데이터들을 쉽게 생성.

1.4.2 SEMMA – 데이터마이닝의 5단계

(Sampling –> Exploration –> Modification –> Modeling –> Assessment)

Sampling(표본추출)

- 추가(Append)
- 데이터 분할(Data Partition)
- 파일 가져오기(File Import)
- 필터링(Filter)
- 입력 데이터(Input Data)
- 병합(Merge)
- 표본추출(Sample)

Exploration(탐색)

- 연관성분석(Association)
- 클러스터링(Cluster)
- DMDB
- 그래프 탐색(Graph Explore)
- 링크 분석(Link Analysis)
- 장바구니 분석(Market Basket)
- 멀티플롯(MultiPlot)
- 경로분석(Path Analysis)
- SOM/Kohonen
- 통계량 탐색(StatExplore)
- 변수 클러스터링(Variable Clustering)
- 변수 선택(Variable Selection)

Modification(수정)

- 제거(Drop)
- 결측값 처리(Impute)
- 대화식 구간 생성(Interactive Binning)
- 주성분분석(Principal Components)
- 값 대체(Replacement)
- 규칙 빌더(Rules Builder)
- 변수 변환(Transform Variables)

Modeling(모형화)

- 자동신경망(AutoNeural)
- 의사결정트리(Decision Tree)
- Dmine 회귀분석(Dmine Regression)
- DM 신경망(DMNeural)
- 앙상블(Ensemble)
- 그래디언트 부스팅(Gradient Boosting)
- LARS
- MBR
- 모델 가져오기(Model Import)
- 신경망(Neural Network)
- 부분최소제곱법(Partial Least Squares)
- 회귀(Regression)
- 규칙추론(Rule Induction)
- TwoStage

Assessment(평가)

- 임계치(Cutoff)
- 의사결정(Decisions)
- 모델비교(Model Comparison)
- 스코어(Score)
- 세그먼트 프로파일링(Segment Profile)

Utility

- 컨트롤 포인트(Control Point)
- 그룹 종료(End Groups)
- Ext Demo
- 메타데이터(Metadata)
- 오픈 소스 통합(Open Source Integration)
- 모델 등록(Register Model)
- 리포트 생성(Reporter)
- SAS 코드(SAS Code)
- 데이터 저장(Save Data)
- 스코어 코드 내보내기(Score Code Export)
- 그룹 시작(Start Groups)

❖ 데이터마이닝에서 흔히 발생하는 문제점

- 장기적이고 구체적인 계획의 부족
- 데이터에 대한 준비 부족
- 시간차이 문제
- 적용상의 문제
- 부서 및 프로젝트들 간의 비협조체제

❖ 데이터마이닝과 통계학의 차이점

- 실험계획법이나 샘플링에서는 사전계획에 따라 자료수집

  데이터 마이닝에서는 대용량의 관측 가능한 자료가 비계획적으로 수집, 관측되어 변수 사이의 관계를 왜곡하는 경우도 있음(교락효과).

- 통계학에서는 표본중심 -> 추정량, 모형구축, p값등이 주요관심사

  데이터마이닝은 모집단중심 ->자료에 대한 탐색, 분석이 중요 (컴퓨터중심).

  → 미래에 대한 예측 중심, 모형구축