데이터 마이닝

2016.08

충북대학교 조 완섭

목차

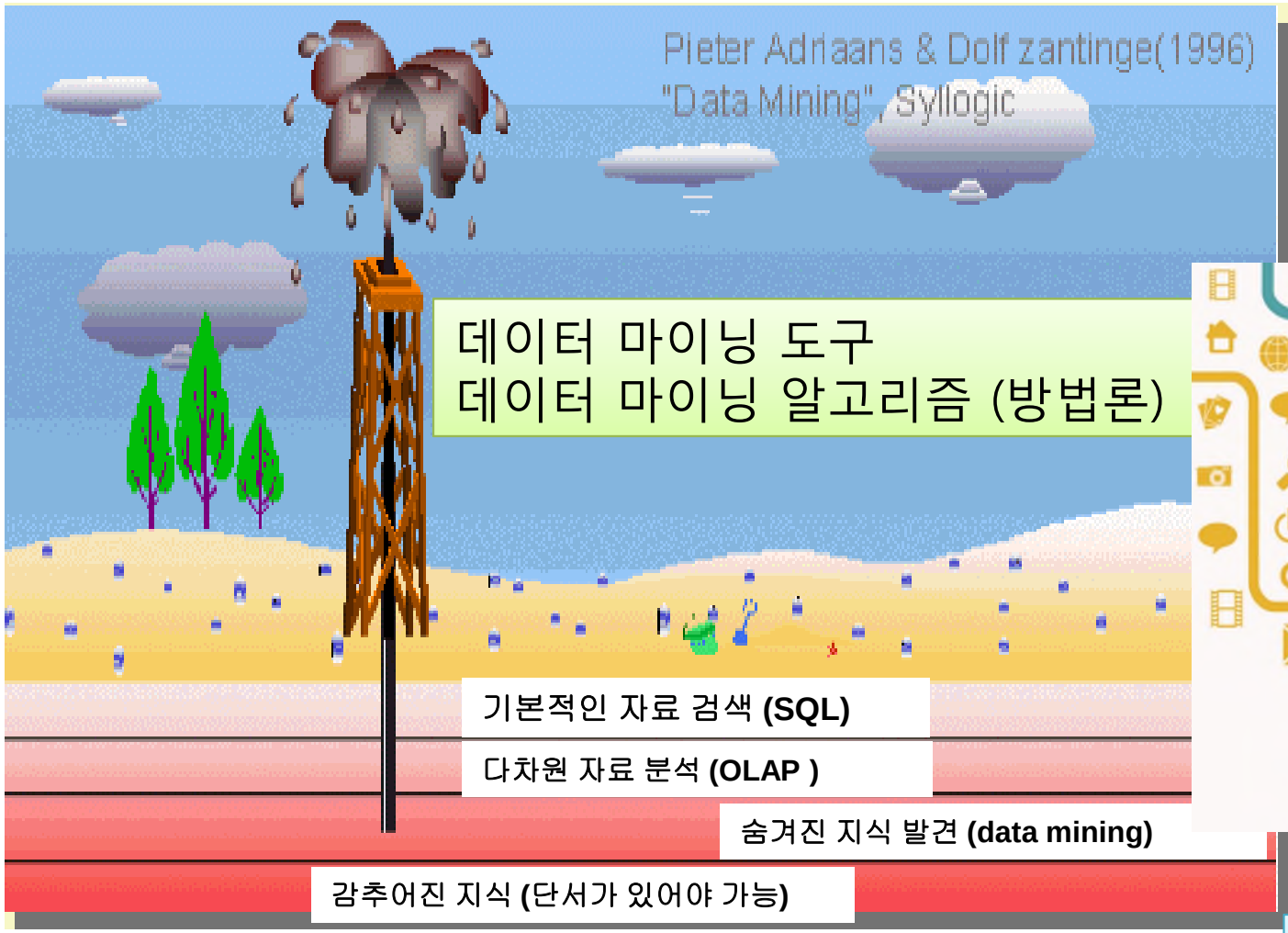
- 데이터 마이닝의 개관
- 연관 규칙
- 분류
- 군집화
- 데이터 마이닝의 다른 문제들
- 데이터 마이닝의 응용들
- 데이터 마이닝 도구

개요

- 데이터 마이닝 (DM)
  - 대용량 데이터로부터 패턴과 규칙 형태의 새로운 지식을 발견하는 작업
  - 데이터 마이닝 결과가 실질적으로 유용하려면 대용량의 파일들이나 데이터베이스에 대하여 마이닝 과정이 수행되어야 하며, DBMS와의 통합이 필요함
  - 여기서는 인공지능, 통계학, 신경망, 유전자 알고리즘 등 다양한 데이터 마이닝 분야를 깊이 다루는 대신 현재 데이터 마이닝 분야의 상황을 간략히 살펴봄
- 데이터 마이닝 분야의 장래
  - Gartner Report 등에서 데이터 마이닝을 가까운 장래의 가장 유망한 기술중 하나로 주목하고 있음

개요

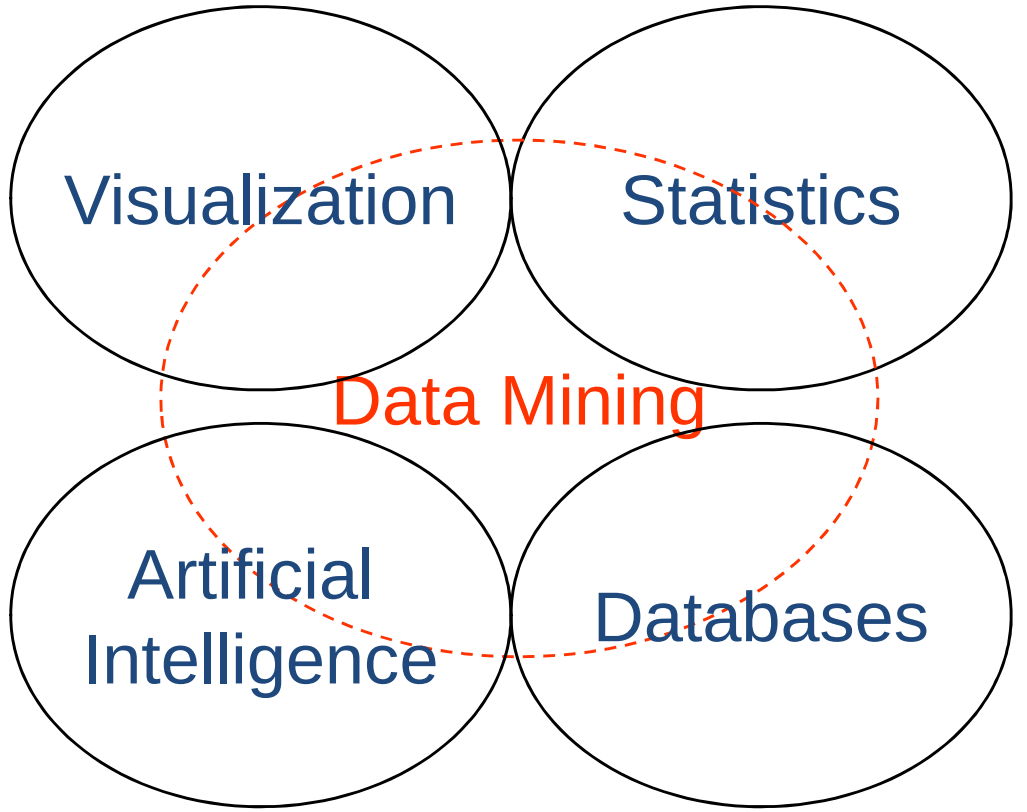
Knowledge



기존 데이터
작고, 정형화된, 느린 데이터

개요

- 데이터 마이닝과 관련학문



개요

- 데이터 마이닝 성공요인
  - 명확한 비즈니스 문제에 대한 인식과 정의
    - 문제 이해
    - 목적의 명확한 정의
  - 충분한 양질의 데이터
    - DW + 그 외 필요한 자료
    - Garbage in garbage out
  - Good mining tools
  - 숙련된 마이너
  - 원활한 조직간 협력과 전사적 지원

OLAP 과 Data Mining

- 데이터 마이닝과 데이터웨어하우징
  - 데이터 마이닝은 데이터 웨어하우스 내의 가공된 데이터나 메타 데이터 혹은 단순 질의에 의해 발견할 수 없는 의미 있는 새로운 지식의 발견을 도와줌
  - 데이터 마이닝 응용들은 DW 설계 초기 단계에서 중요하게 고려되어야 하며, 데이터 마이닝 도구들도 데이터 웨어하우스와의 연계 사용을 감안하여 개발되는 것이 바람직
  - 수십 테라 바이트의 대용량 데이터베이스에서 데이터 마이닝 응용의 성공적인 수행 여부는 DW의 구축에 크게 의존

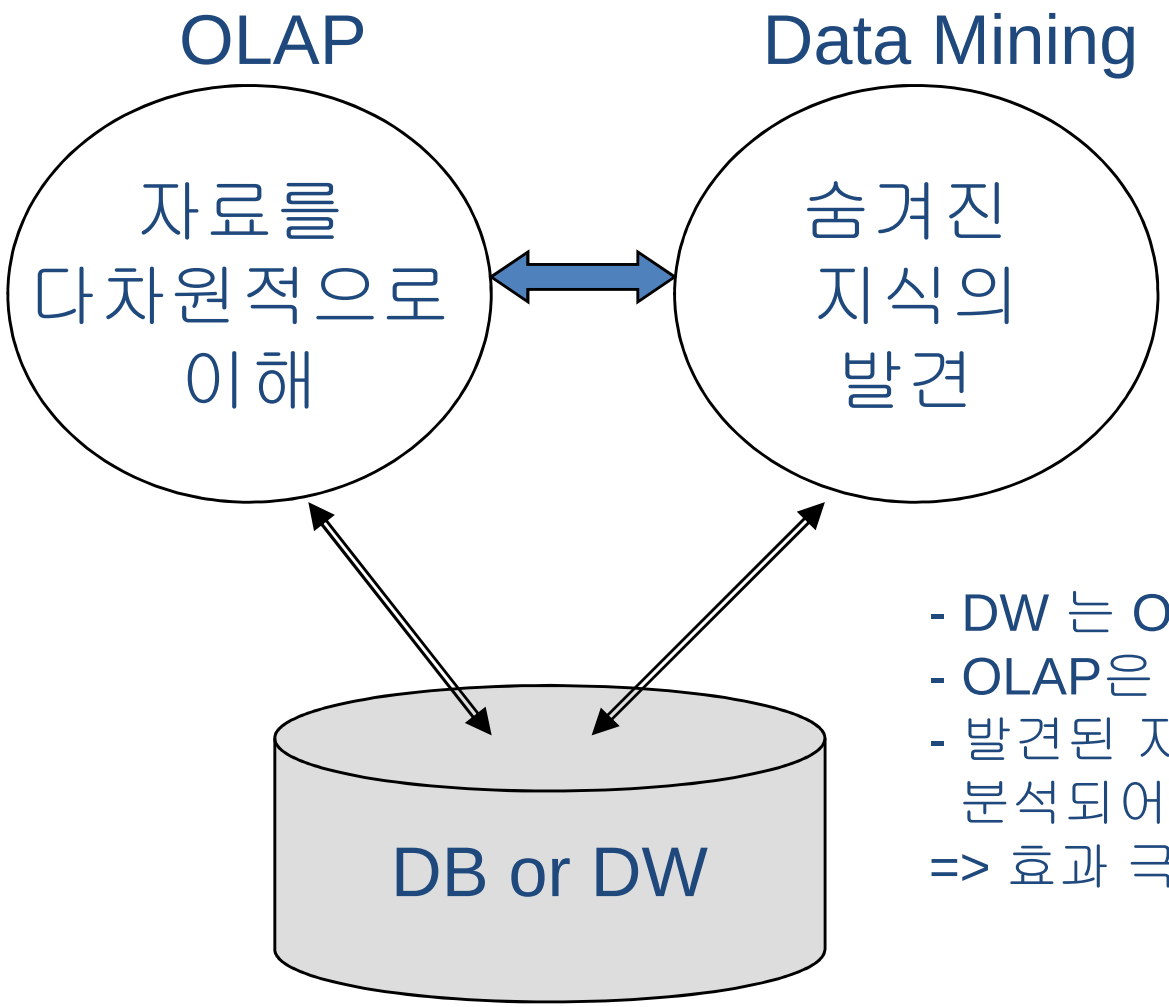
OLAP 과 Data Mining

- OLAP & DM의 차이

OLAP	Data Mining
▪주어진 자료를 미리 정해진 다차원으로 분석/요약	▪ 예상치 못한 지식의 발견
▪ What?-why 에 대한 대답	▪ Why?에 대한 대답
▪ 정해진 가설의 확인	▪ 예측 및 룰의 발견
▪ 데이터에 대한 기본적인 이해 증진	▪ 의사결정의 고급정보 제공
▪ 데이터 마이닝의 효과 극대화	▪ OLAP의 차원 정보로 feedback
Mailing 결과 지역별/연령별 응답율은 ?	Mailing을 받고 응답할 것 같은 고객의 프로파일은 ?
기존 고객은 신상품 중 어느 것을 주로 구매했나 ?	신상품을 구매할 것 같은 고객은 ?
지난해 수익성이 높은 고객 top 10은 ?	어떤 특징을 지닌 고객들이 수익성이 가장 높은가 ?
이탈한 고객의 지역별/분기별 분포는 ?	이탈 고객의 특성은 ?
재무상태가 좋지 않은 고객은 누구인가 ?	재무 상태가 좋지 않은 고객의 특성은 ?

OLAP 과 Data Mining

- 상호 보완적 관계



- DW 는 OLAP/DM의 기반 제공
- OLAP은 DM에 유용한 정보제공 (단서)
- 발견된 지식은 DW & DB에 저장,
  분석되어 상호 보완적 운영

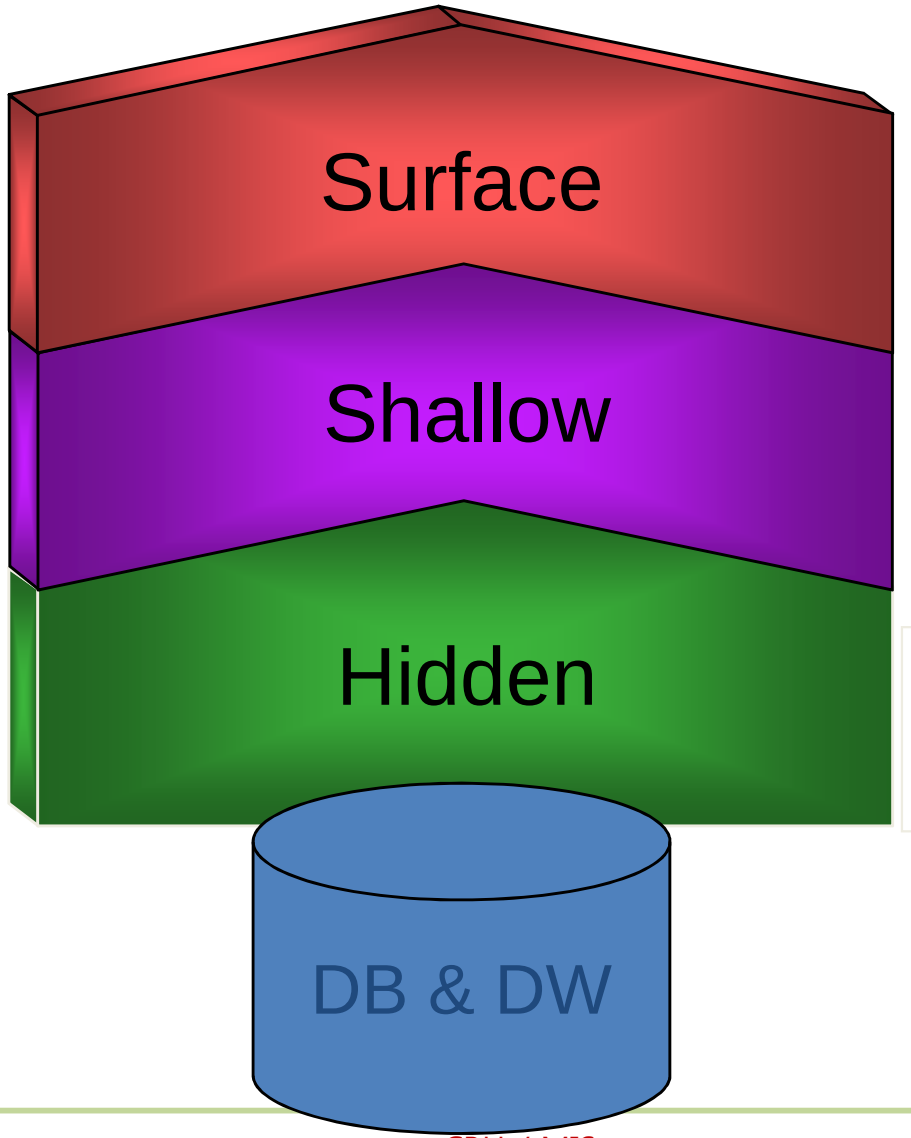
=> 효과 극대화

OLAP 과 Data Mining

- 비교

Top-Down Methodology

Bottom-Up Methodology



Analytical Tools Used

SQL (Structured Query Language) for simple queries and reporting

Statistical & OLAP for summaries, analysis, & forecasting

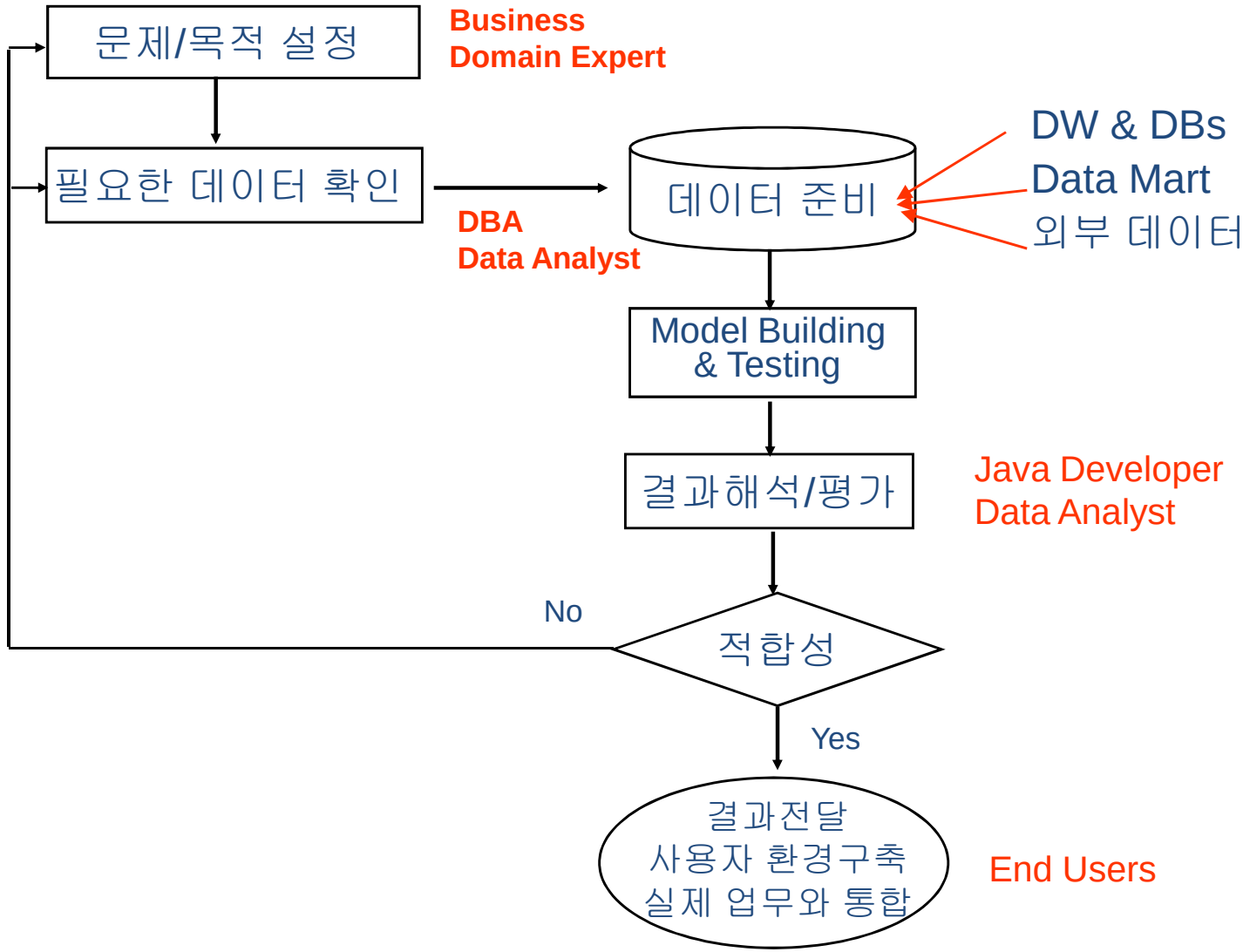
Data Mining for classification, clustering and predictions

Data Mining 프로세스

- 지식 발견 과정으로서의 데이터 마이닝
  - 데이터베이스 내에서 단순검색과 집계를 넘어 지식 발견은 중요한 의미를 가짐 => 인공지능
  - 지식 발견의 단계
    - 데이터 추출(data selection)
    - 데이터 정제(data cleansing)
    - 데이터 내용 강화(data enrichment)
    - 데이터 변형(data transformation) 또는 인코딩(encoding)
    - 데이터 검색, 집계, 다차원분석
    - **데이터 마이닝(data mining)**
    - 보고서 작성(reporting) 단계
  - 데이터 마이닝에서는 데이터로부터 연관규칙, 연속패턴, 분류트리 등을 발견함

Data Mining 프로세스

- 데이터 마이닝 절차 (Cycle)



Data Mining 프로세스

- Ad-hoc vs. Repeatable approaches
  - Data mining tools: ad hoc data mining
    - One or two (data analyst) users
    - Extensive overhead of extracting and preparing data for each data mining exercise
    - Data stays in the files or databases
    - One-time results; not a repeatable process
  - Data mining infrastructure: repeatable data mining
    - Many users throughout the organization
    - Data stays in the databases or DW's
    - Automatically sift through data to find new business intelligence
    - Enable applications with predictions and insights
    - Data mining benefits for non-expert(s)

마이닝의 목표와 종류

- 마이닝과 지식 발견의 목표
  - 데이터 마이닝의 목표는 **예측, 식별, 분류, 최적화의** 네가지로 분류됨
  - **예측**에서는 향후 발생할 사건을 예측함 – 구매 고객 예측, 판매량 예측 등
  - **식별**에서는 사건(event) 및 활동(activity)의 존재를 식별하는데 이용되는 패턴을 발견함 – 해커의 활동 패턴 발견, 유전자 패턴 식별 등
  - **분류**에서는 데이터를 클래스 혹은 카테고리로 분할함 – 쇼핑몰에서 고객의 분류
  - **최적화**에서는 주어진 제약 조건 하에서 시간과 공간, 자금과 재료 등과 같은 제한된 자원을 최적으로 사용하여 이익을 최대화하는데 있음

마이닝의 목표와 종류

- 마이닝 기법의 종류와 응용분야

기능	기법	적용분야
연관분석 association	Association rules generations	장바구니 분석
분류 classification	의사결정나무 사례기반 추론 신경망 판별분석/로짓분석	타겟마케팅 신용평가 질병 진단 등
군집분석 clustering	신경망 K-means algorithm 의사결정나무	시장세분화 Web 구조 개선
순차패턴 Seq. patterns	Sequential pattern analysis	시간 개념을 이용한 장바구니 분석
예측 forecasting	시계열분석 회귀분석 사례기반추론(case-based reasoning) 신경망 등	주가/환율 예측 수요예측 재고 및 품질관리 등

연관분석

- 연관분석(association analysis)
  - 한 데이터와 다른 데이터 사이의 관련성이 있음을 찾는 규칙
  - 연관규칙은 X⇒Y의 형태로 표현됨
    - 구매 부문에서 X⇒Y의 의미는 만일 한 고객이 X를 구매하면, Y도 함께 구매할 가능성이 있음을 의미함
  - 지지도(support) : 전체 트랜잭션 중에서 X∪Y 항목들이 함께 나타나는 트랜잭션들의 비율
  - 신뢰도(confidence) : X를 포함하는 트랜잭션 중에서 Y까지 포함하는 트랜잭션의 비율

Transaction-id	Time	Items-Brought
101	6:35	milk, bread, juice
792	7:38	milk, juice
1130	8:05	milk, eggs
1735	8:40	bread, cookies, coffee

Milk⇒Juice는 50%의 지지도와 66.7%의 신뢰도를 가짐
Bread⇒Juice는 25%의 지지도와 50% 의 신뢰도를 가짐

연관분석

- 연관분석 과정 탐사하는 2가지 단계
  - 최소 지지도 이상인 항목 집합(빈발항목집합, frequently itemset)을 생성해 나가는 방식으로 연관규칙을 찾음

transaction database

1: $\{a, d, e\}$
2: $\{b, c, d\}$
3: $\{a, c, e\}$
4: $\{a, c, d, e\}$
5: $\{a, e\}$
6: $\{a, c, d\}$
7: $\{b, c\}$
8: $\{a, c, d, e\}$
9: $\{b, c, e\}$
10: $\{a, d, e\}$

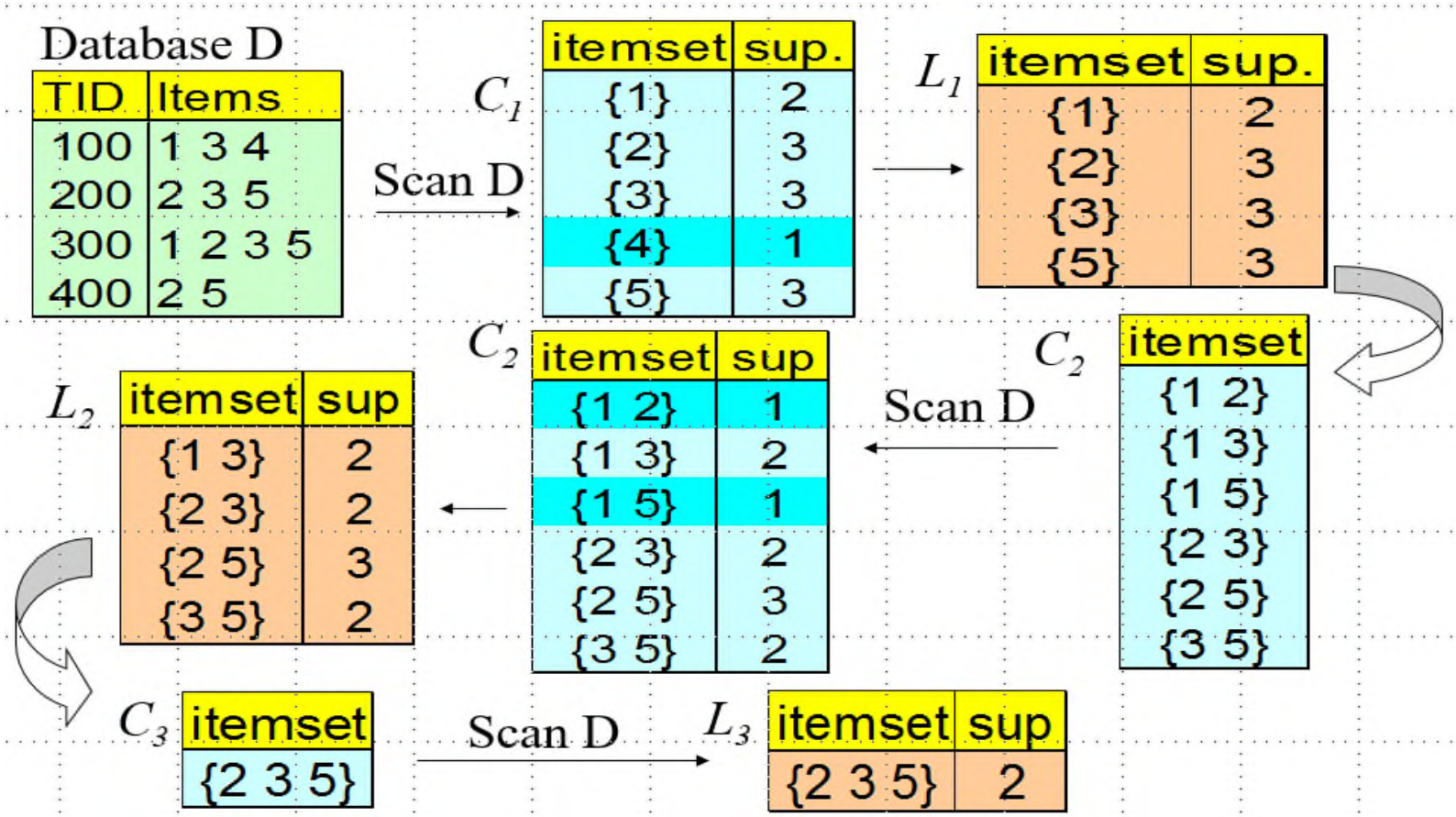
frequent item sets

0 items	1 item	2 items	3 items
$\emptyset$: 10	$\{a\}$: 7	$\{a, c\}$: 4	$\{a, c, d\}$: 3
	$\{b\}$: 3	$\{a, d\}$: 5	$\{a, c, e\}$: 3
	$\{c\}$: 7	$\{a, e\}$: 6	$\{a, d, e\}$: 4
	$\{d\}$: 6	$\{b, c\}$: 3	
	$\{e\}$: 7	$\{c, d\}$: 4	
		$\{c, e\}$: 4	
		$\{d, e\}$: 4	

- In this example, the minimum support is $s_{\min} = 3$ or $\sigma_{\min} = 0.3 = 30\%$.
- There are $2^5 = 32$ possible item sets over $B = \{a, b, c, d, e\}$.
- There are 16 frequent item sets (but only 10 transactions).

연관분석

- 항목의 개수가 많은 경우에 문제점
  - 항목의 개수가 m 개이면 서로 다른 항목집합의 수는 2^m 개이며, 따라서 계산 오버헤드가 심각해짐 (1천개 항목 => 2^{1000})
  - Apriori 알고리즘, 샘플링 알고리즘, 빈발-패턴 트리 알고리즘, 분할 알고리즘 등이 제안되었으며, 대규모 항목집합에서 성능향상에 초점을 맞춤



연관분석

- 계층구조들간의 연관규칙
  - 응용 분야의 특성상 항목집합을 계층 형태로 구분하는 것이 자연스럽다면 계층 내의 연관 규칙과 함께 계층간에 존재하는 연관규칙을 발견하는 것이 특별한 의미를 가짐 (다음 slide)

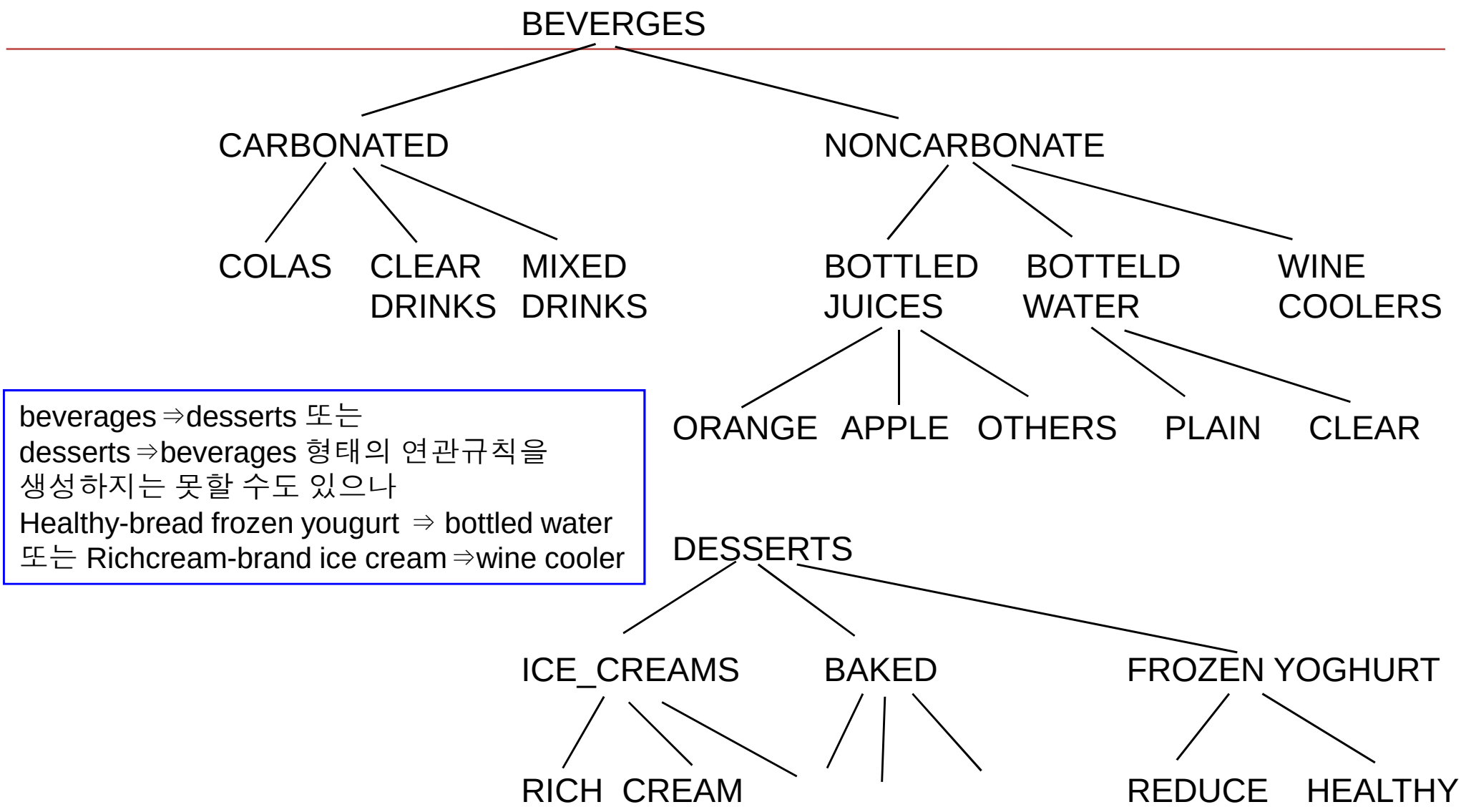


그림. 슈퍼마켓에서 항목들의 계층 구조

연관분석

- 다차원 연관성
  - 지금까지 소개한 연관규칙은 단지 하나의 차원 (속성)만을 포함하지만 실제로 두 개 이상의 차원에 대한 연관규칙도 중요함
  - 예
    - 단일차원 연관규칙 : 구입한 물건(우유)=>구입한 물건(주스)
    - 2차원 연관규칙 : 시간(6:30...8:00)=>구입한 물건(우유)
  - 차원들은 범주(예: 구입한 물품)나 양적(예: 시간, 소득) 속성이 될 수 있음
  - 양적 속성은 값들을 겹치지 않는 구간들로 파티션하고, 각 구간에 레이블을 주는 방식이 주로 사용됨
    - 예 : 저소득 (급여<1000만원), 중간소득(1000만원<=급여<5000만원), 고소득(5000만원<=급여)

연관분석

- 부정 연관성
  - 두 항목간에 연관이 없음을 나타내는 규칙
  - 예: 포테이토 칩을 사는 고객중에서 60%는 병에 든 물을 사지 않는다. (여기서 60%는 부정 연관규칙의 신뢰도를 가리킨다.)
  - 발견되는 부정 연관성의 규칙 중에는 유용하지 않은 것이 많을 수 있음
  - 관심있는 부정 연관성을 발견하기 위해서는 도메인 지식을 이용하는 것이 것이 중요함

연관분석

- 연관 규칙을 위한 부가적 고려사항들
  - 대부분 상황에서 항목 집합들의 카디널리티는 매우 크며, 트랜잭션도 많다. 소매업과 통신회사들에서 운영하는 데이터베이스에는 하루에 수천만 개의 트랜잭션이 모아진다.
  - 트랜잭션들은 지리적 위치나 계절과 같은 요인에 민감할 수 있으며, 이것이 샘플링을 더욱 어렵게 만들게 된다.
  - 항목의 분류도 여러 차원에서 이루어질 수 있다. 그러므로 도메인 지식을 가지고 부정 규칙을 발견하는 것이 어려울 수 있다.
  - 데이터의 질(quality)도 변화한다. 따라서 여러 기업으로부터 입력되는 데이터의 중복 뿐 아니라, 데이터 결여와 데이터 오류 및 데이터 불일치 등과 관련된 중요한 문제가 발생할 수 있다.

분류

- 분류의 정의
  - 데이터를 서로 다른 부류 혹은 클래스 (미리 결정되어 있음)로 나누는 모델을 학습하는 과정
  - classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data [*Jiawei Han*]
  - 예를 들어, 신용카드를 신청한 고객들을 “poor risk”, “fair risk”, “good risk”로 나누어 주는 모델의 개발
  - 모델은 대개 **결정트리(decision tree)**나 **규칙들의 집합** 형태로 표시됨
  - 이미 분류되어 있는 훈련 데이터 집합을 사용하여 일단 모델이 구축하고, 그 모델을 이용하여 새로운 데이터를 분류함

분류

- 결정트리
  - 데이터에 대한 분류 규칙들을 트리 형태로 간단히 표현한 것

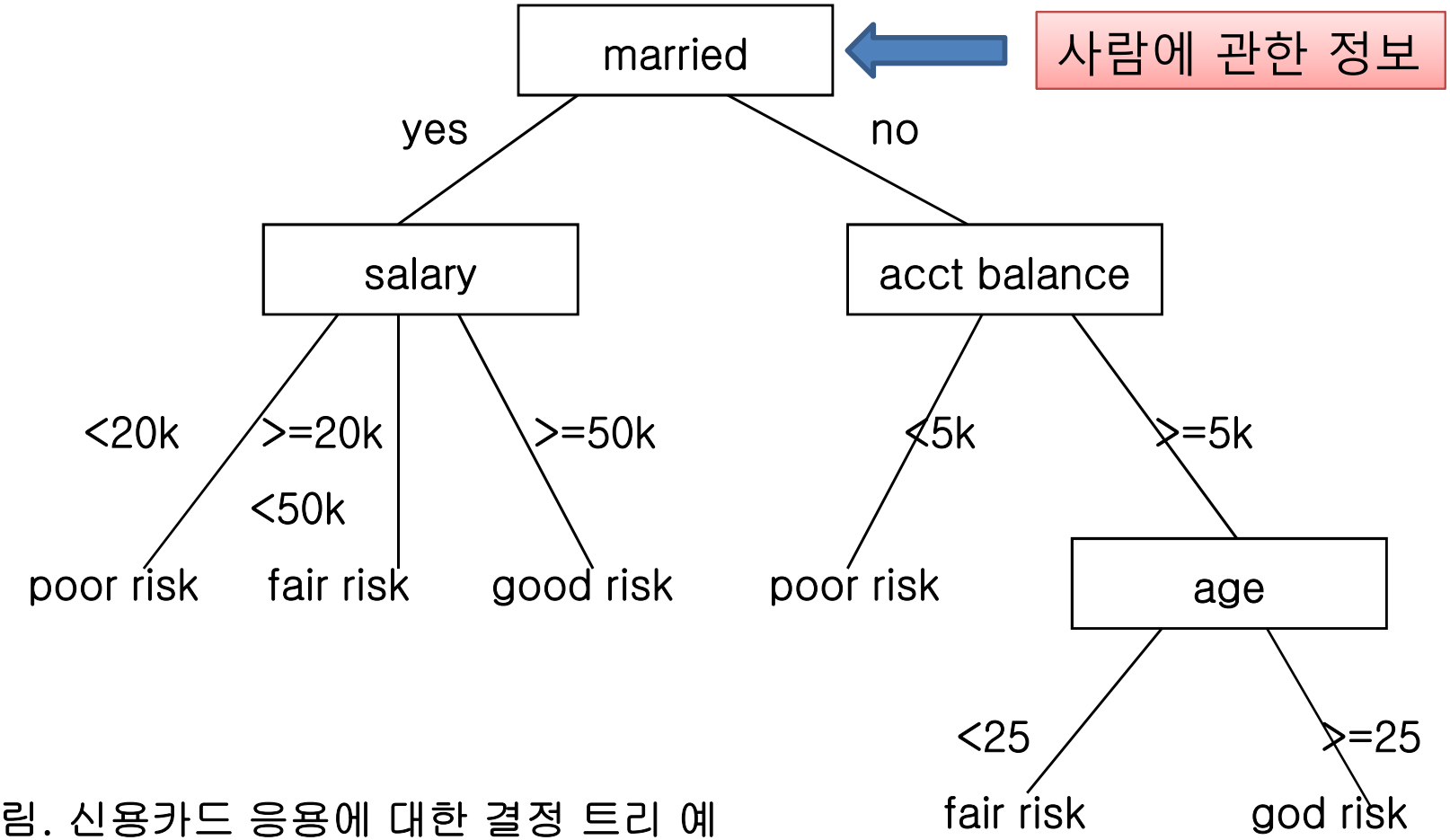
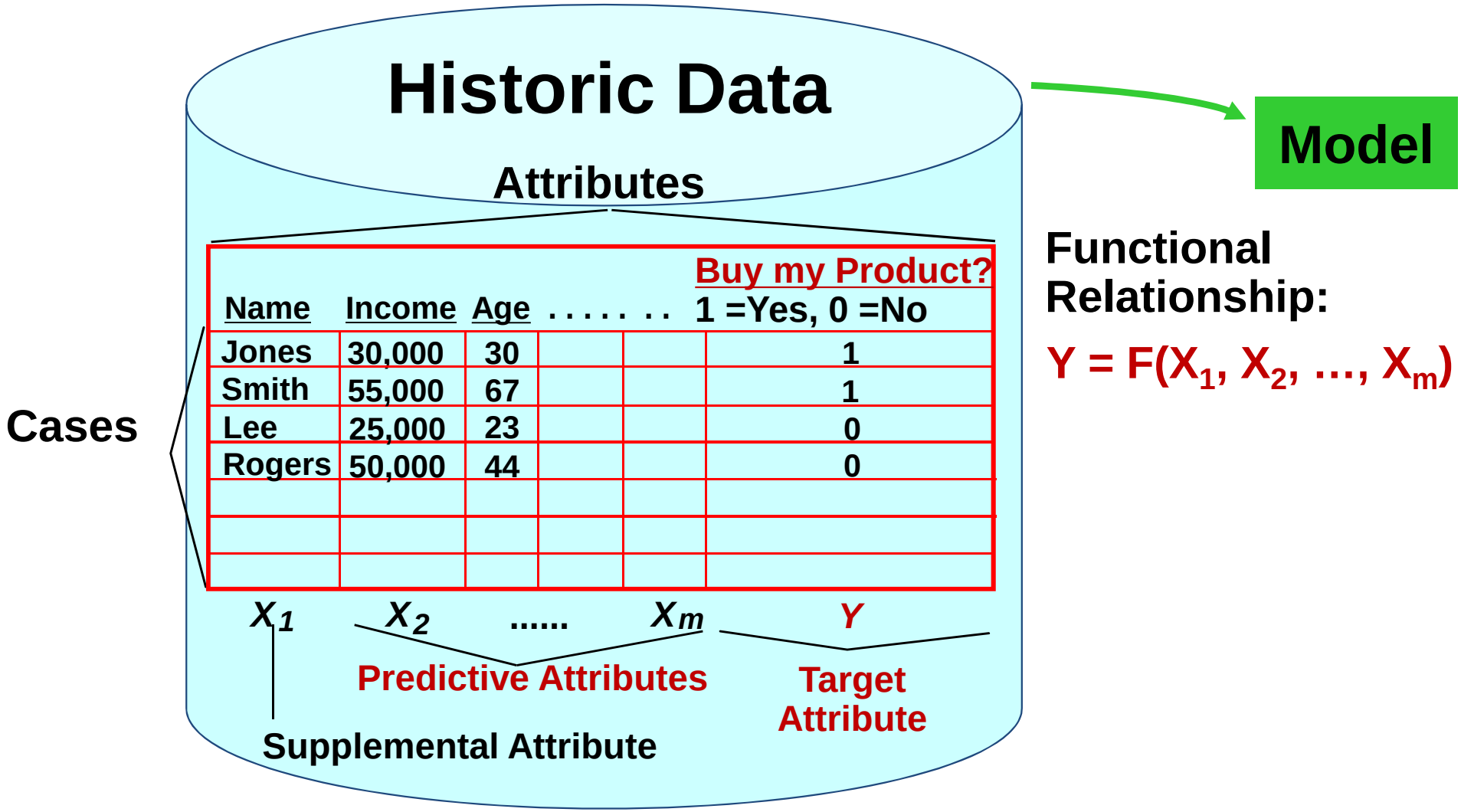


그림. 신용카드 응용에 대한 결정 트리 예

분류

- 훈련단계

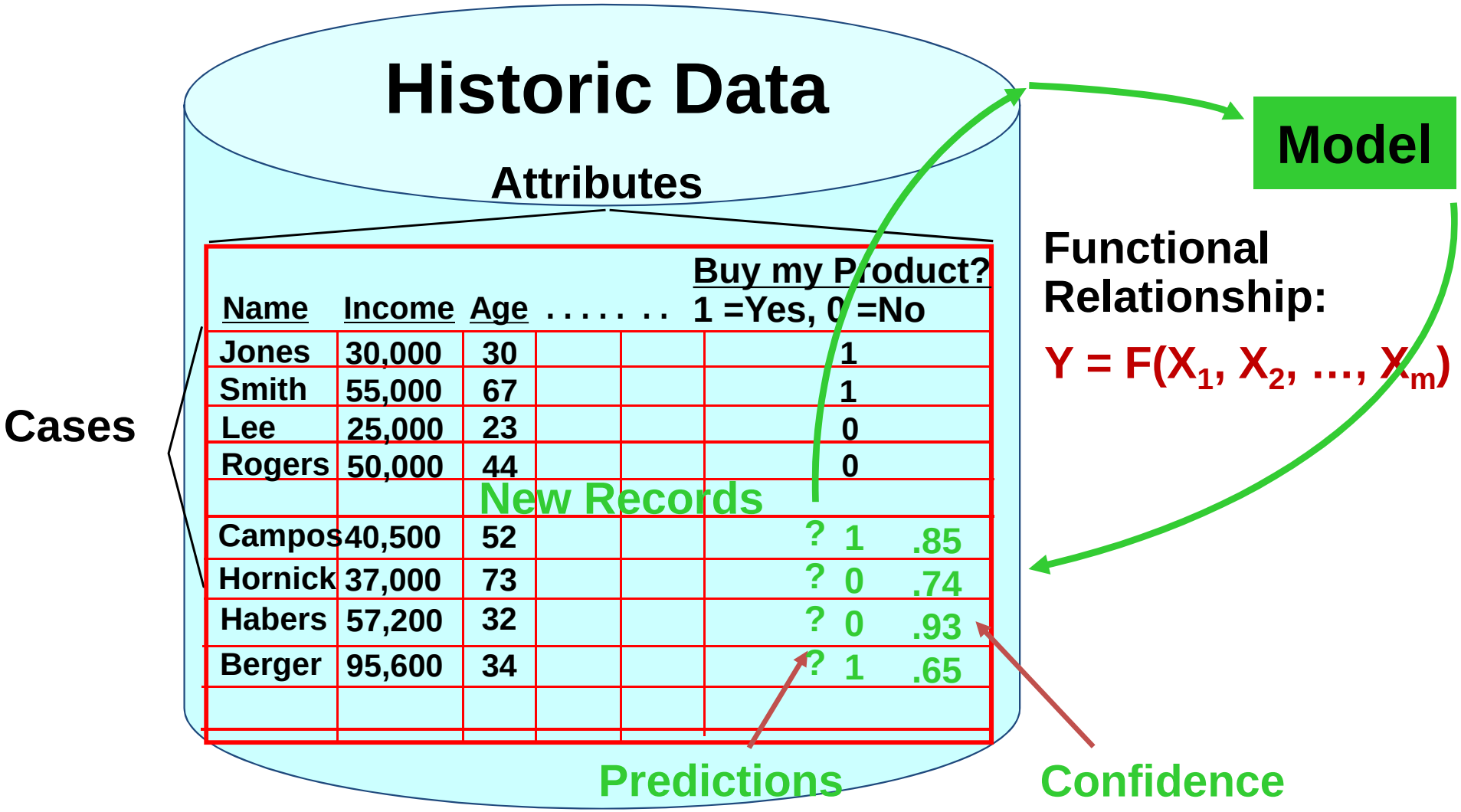
Classification : (1) model construction



분류

- 실전단계

(2) Classification of new data



분류 – 다른 예제 (2)

The weather data with ID code

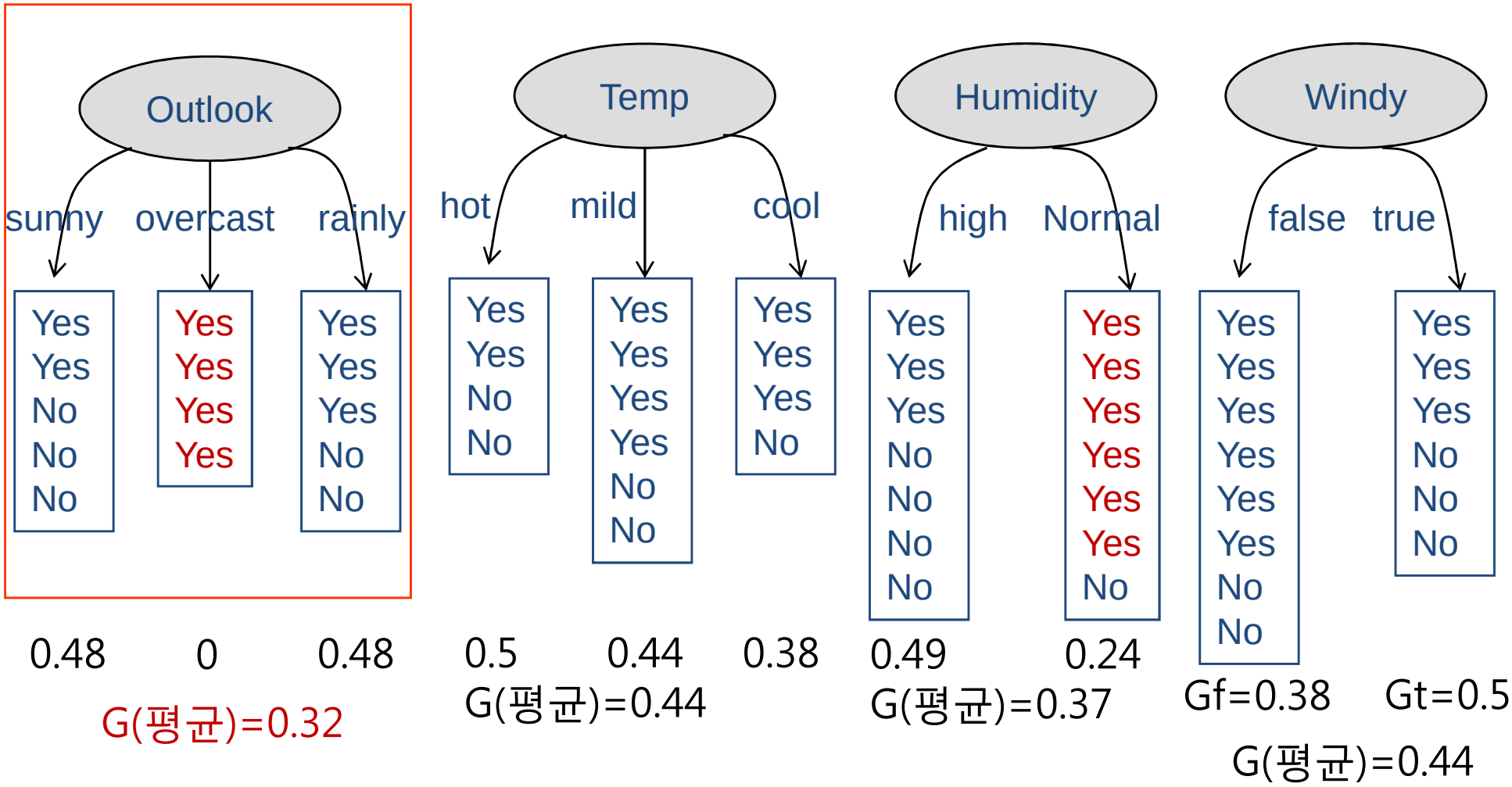
IDCode	Outlook	Temp	Humidity	Windy	Play
A	Sunny	Hot	High	F	No
B	Sunny	Hot	High	T	No
C	Overcast	Hot	High	F	Yes
D	Rainly	Mild	High	F	Yes
E	Rainly	Cool	Normal	F	Yes
F	Rainly	Cool	Normal	T	No
G	Overcast	Cool	Normal	T	Yes
H	Sunny	Mild	High	F	No
I	Sunny	Cool	Normal	F	Yes
J	Rainly	Mild	Normal	F	Yes
K	Sunny	Mild	Normal	T	Yes
L	Overcast	Mild	High	T	Yes
M	Overcast	Hot	Normal	F	Yes
N	Rainly	Mild	High	T	No

Input data
- Outlook
- Temp
- Humidity
- Windy

Target Attribute
- Play

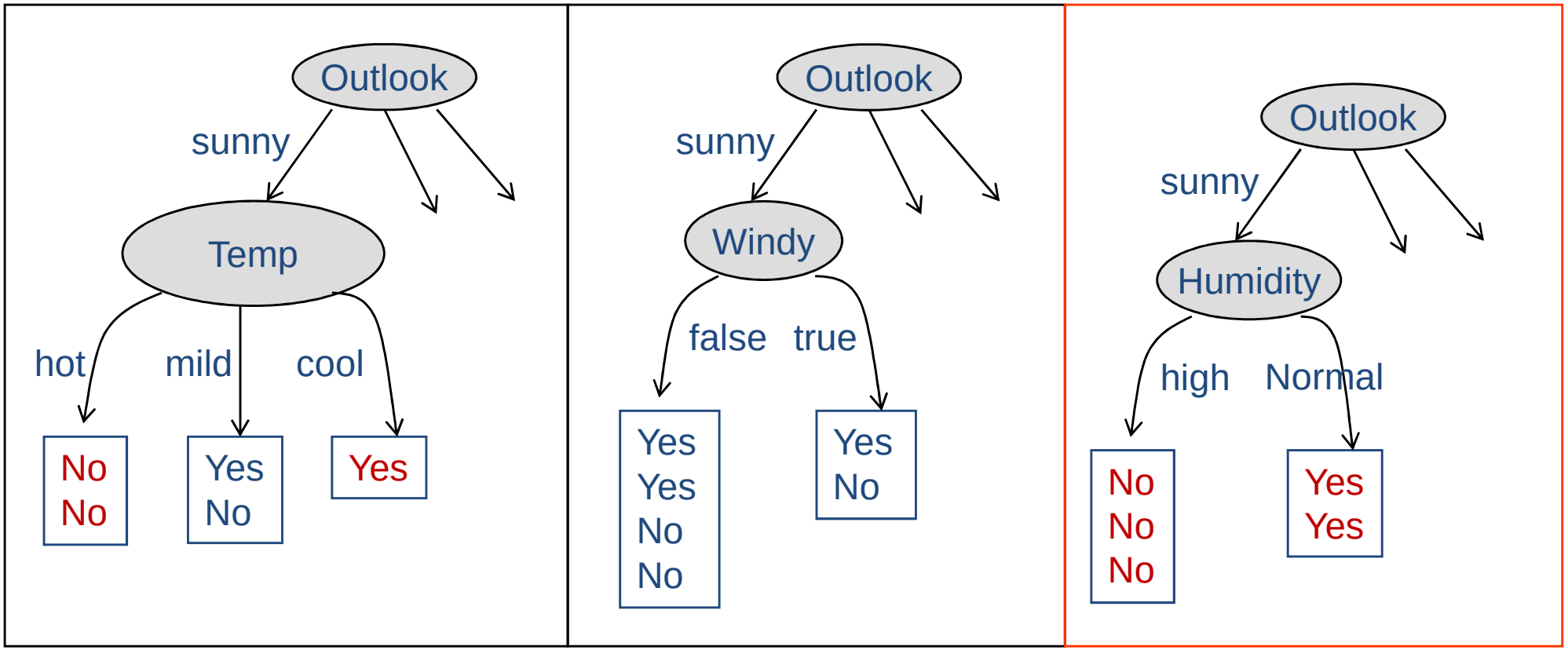
분류 – 다른 예제 (2)

- Which attribute to select ?



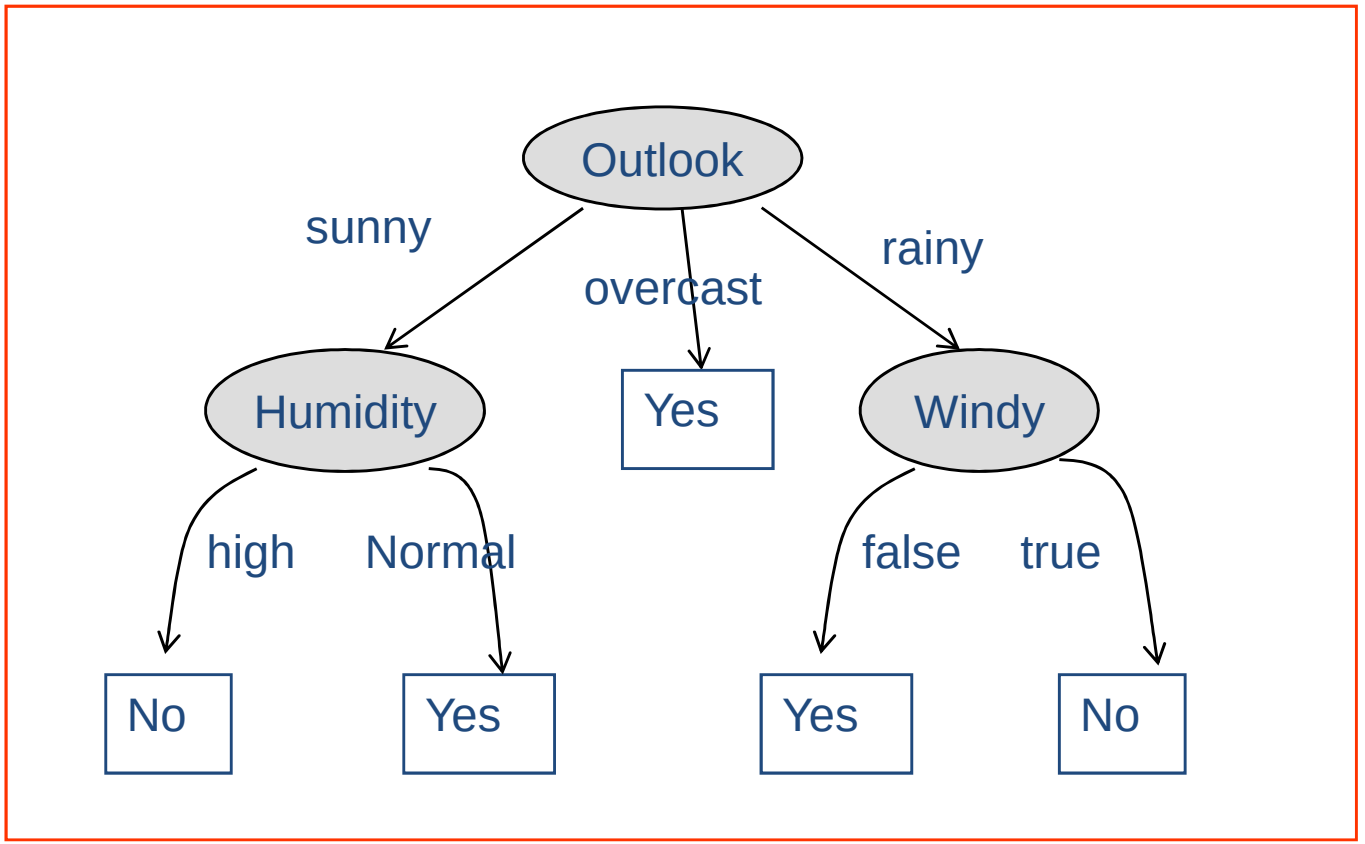
분류 – 다른 예제 (2)

- Which attribute to select ?



분류 – 다른 예제 (2)

- Which attribute to select ?



분류

- 결정트리 생성법
  - 자식노드로 분할할 때 각 노드에 포함된 데이터의 동질성이 높게 (다양성이 최소화되도록) 하는 것이 바람직함
  - 세가지 방법이 사용됨
    - 피어슨 카이제곱 검정에 대한 P값(Chi-square)
    - 지니지수(Gini Index)
    - 엔트로피(Entropy)

분류

- **지니계수 :** 데이터셋에서 임의로 두 원소를 복원추출할 때 서로다를 확률
  - 원소가 동질할수록 **0**에 가깝고, 이질적일수록 **1**에 가까운 값을 갖는 척도

Set1

A, B, C, A, C, C, A, D

$$G(Set1) = 1 - \left(\frac{\text{원소}A\text{의 개수}}{\text{전체 원소갯수}}\right)^2 - \left(\frac{\text{원소}B\text{의 개수}}{\text{전체 원소갯수}}\right)^2 - \left(\frac{\text{원소}C\text{의 개수}}{\text{전체 원소갯수}}\right)^2 - \left(\frac{\text{원소}D\text{의 개수}}{\text{전체 원소갯수}}\right)^2$$
$$= 1 - \left(\frac{3}{8}\right)^2 - \left(\frac{1}{8}\right)^2 - \left(\frac{3}{8}\right)^2 - \left(\frac{1}{8}\right)^2 = 0.69$$

Set2

A, A, A, A, B, B, B, B

$$G(Set2) = 1 - \left(\frac{\text{원소}A\text{의 개수}}{\text{전체 원소갯수}}\right)^2 - \left(\frac{\text{원소}B\text{의 개수}}{\text{전체 원소갯수}}\right)^2$$
$$= 1 - \left(\frac{4}{8}\right)^2 - \left(\frac{4}{8}\right)^2 = 0.5$$

Set3

A, A, A, A, A, A, A, A

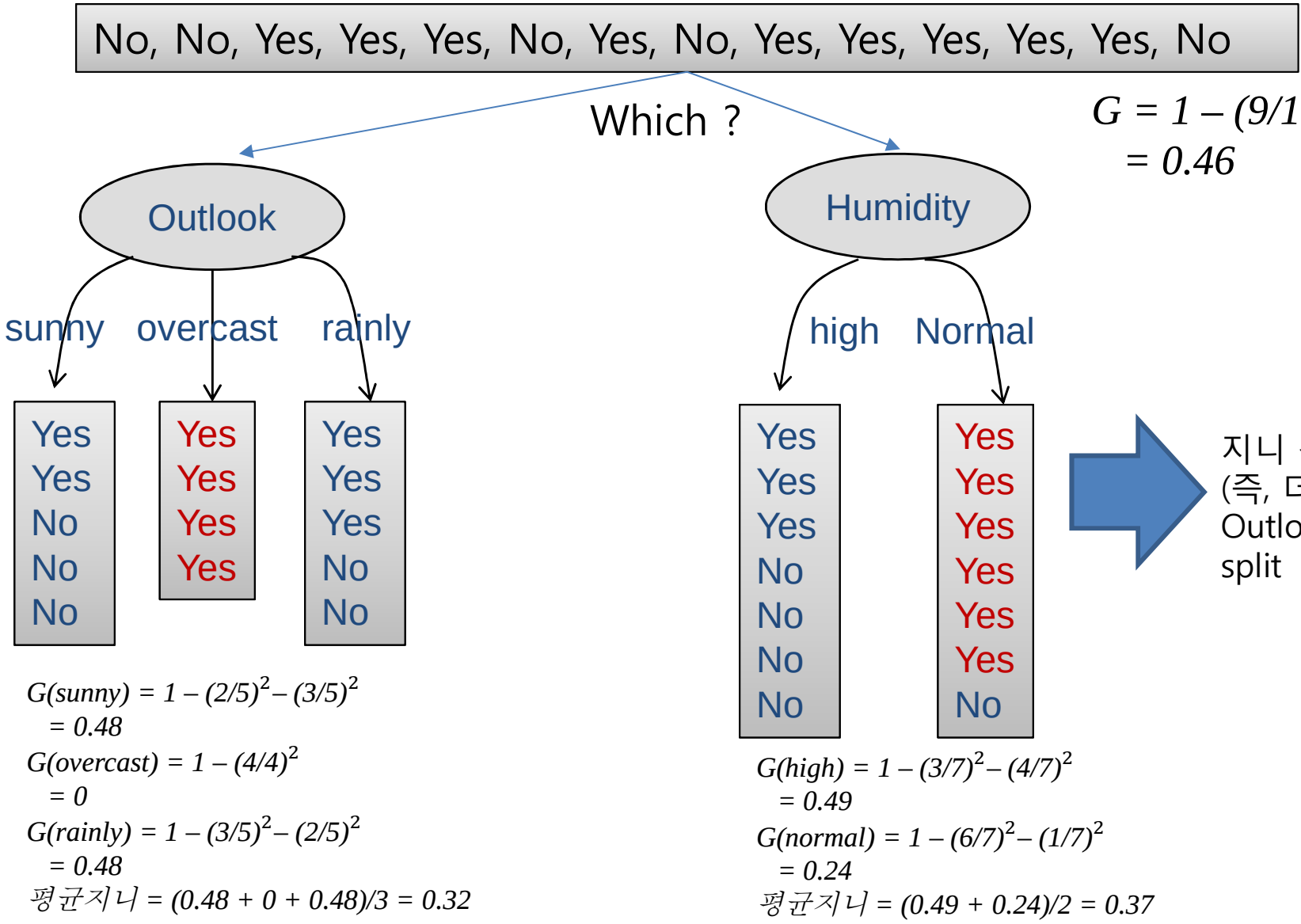
$$G(Set3) = 1 - \left(\frac{\text{원소}A\text{의 개수}}{\text{전체 원소갯수}}\right)^2$$
$$= 1 - \left(\frac{8}{8}\right)^2 = 0$$

Set4

A, B, C, D, E, F, G, H

$$G(Set4) = 1 - 8\left(\frac{1}{8}\right)^2 = 0.88$$

분류



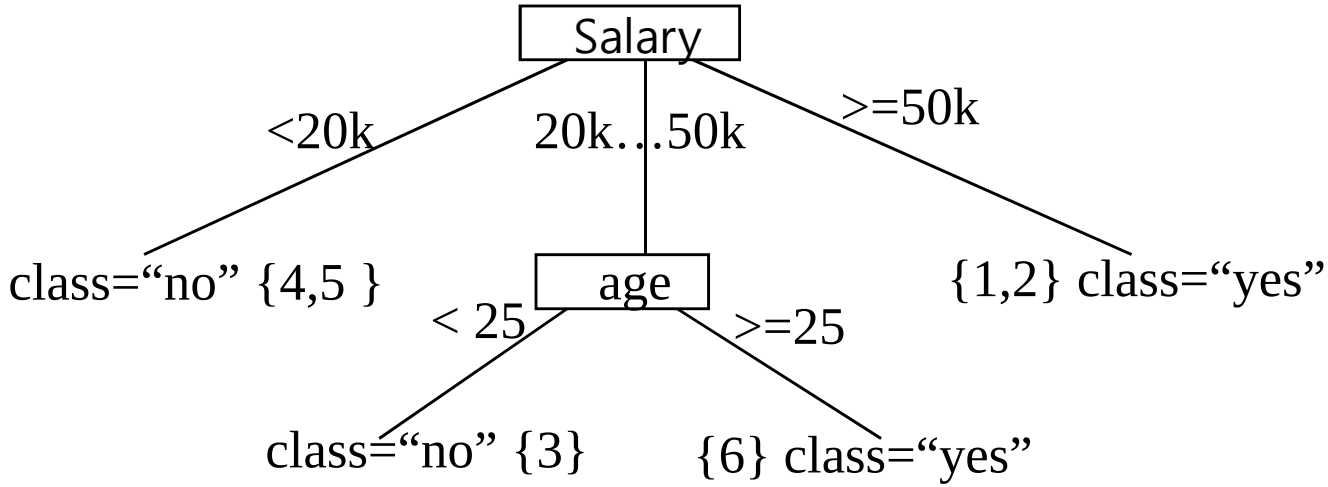
$G = 1 - (9/14)^2 - (5/14)^2 = 0.46$

지니 감소량이 더 많은 (즉, 더 동질하게 분할하는) Outlook를 기준으로 split

분류

- 또 다른 예제

Rid	Married	Salary	Acct balance	Age	Loanworthy (class)
1	no	>=50k	<5k	>=25	yes
2	yes	>=50k	>=5k	>=25	yes
3	yes	20k…50k	<5k	<25	no
4	no	<20k	>=5k	<25	no
5	no	<20k	<5k	>=25	no
6	yes	20k…50k	>=5k	>=25	yes



군집화

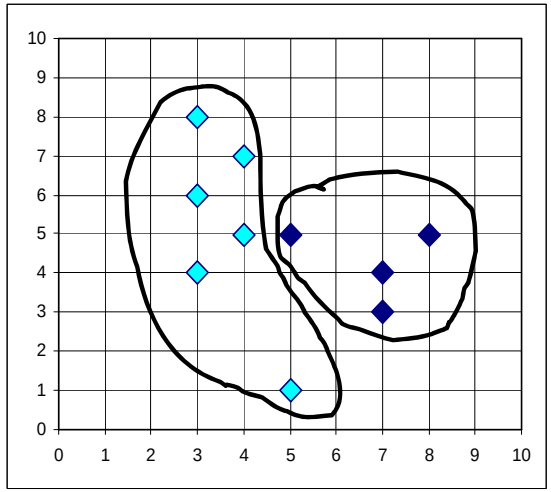
- 군집화
  - 훈련 샘플을 갖지 않는 데이터의 분할에 사용되는 자율적 학습 (unsupervised learning) 형태
  - 유사한 레코드들은 같은 그룹에 배치하고 서로 다른 레코드들은 다른 그룹에 배치시킴; 그룹들 간에는 서로 겹치지 않음
  - 수치데이터일 때, 거리에 바탕을 둔 유사성 함수를 사용해서 유사성을 측정 함
  - K-평균 알고리즘(k-Means algorithm)이 주로 사용 됨

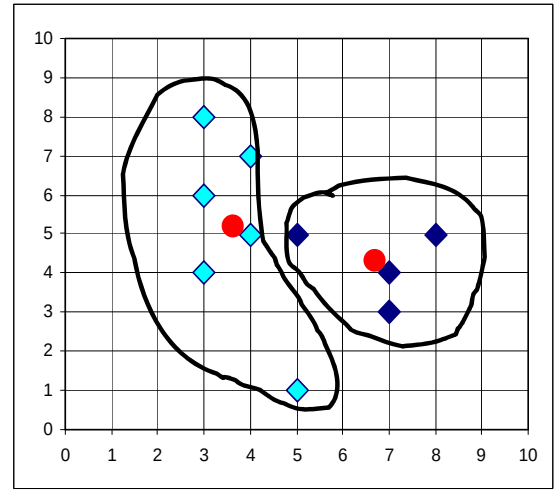
군집화

- K-평균 알고리즘
  - 원하는 클러스터 개수 k를 임의로 선택
  - K개 클러스터를 위한 중간값(means)으로 임의의 k개 레코드 선택
  - 모든 레코드들을 중간값과 레코드의 간격을 기반으로 주어진 클러스터에 배치
  - 각 클러스터의 중간값이 재계산 됨
  - 다시 각 레코드를 조사하여 중간값이 가장 가까운 클러스터에 레코드를 배치
  - 레코드들의 변동이 더 이상 이루어지지 않을 때까지 반복 수행

군집화 : K-평균 알고리즘

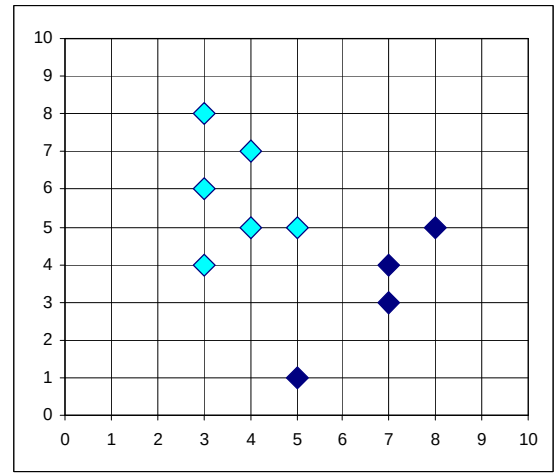
1단계.
레코드를 원하는
클러스터 개수
k=2 로 분할

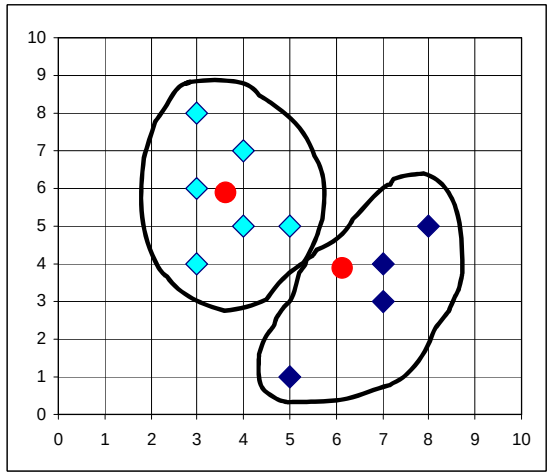




2단계.
현재 분할된
클러스터의
중간값을 계산

3단계.
각 레코드를
중간값이
가장 가까운
클러스터에
재 할당





4단계.
(2단계 반복 수행)
새롭게 분할된
클러스터의
중간값을 재계산

순차패턴

- 순차패턴의 발견
  - 항목집합들이 연속적으로 발생하는 경우, 이들로부터 빈번하게 발생하는 패턴을 발견

    예: 시장 바구니 트랜잭션에서 {milk, bread, juice}, {bread, eggs}, {cookies, milk, coffee}는 한 고객이 상점을 3번 방문하여 구매한 항목집합들의 시퀀스(sequence) 임
  - 여러 고객의 구매 시퀀스들에서 빈번하게 나타나는 부분 시퀀스들을 찾으면 이로부터 고객들의 구매 패턴을 예측할 수 있음

비디오점의 대여 기록 **data**

고객번호	구매기록
1	{겨울연가} => {아폴로13, 캐스트웨이}
2	{겨울연가} => {아폴로13,공동경비구역}
3	{러브레터} => {시월이야기, 동감}{시월애}
4	{겨울연가} => {캐스트웨이}

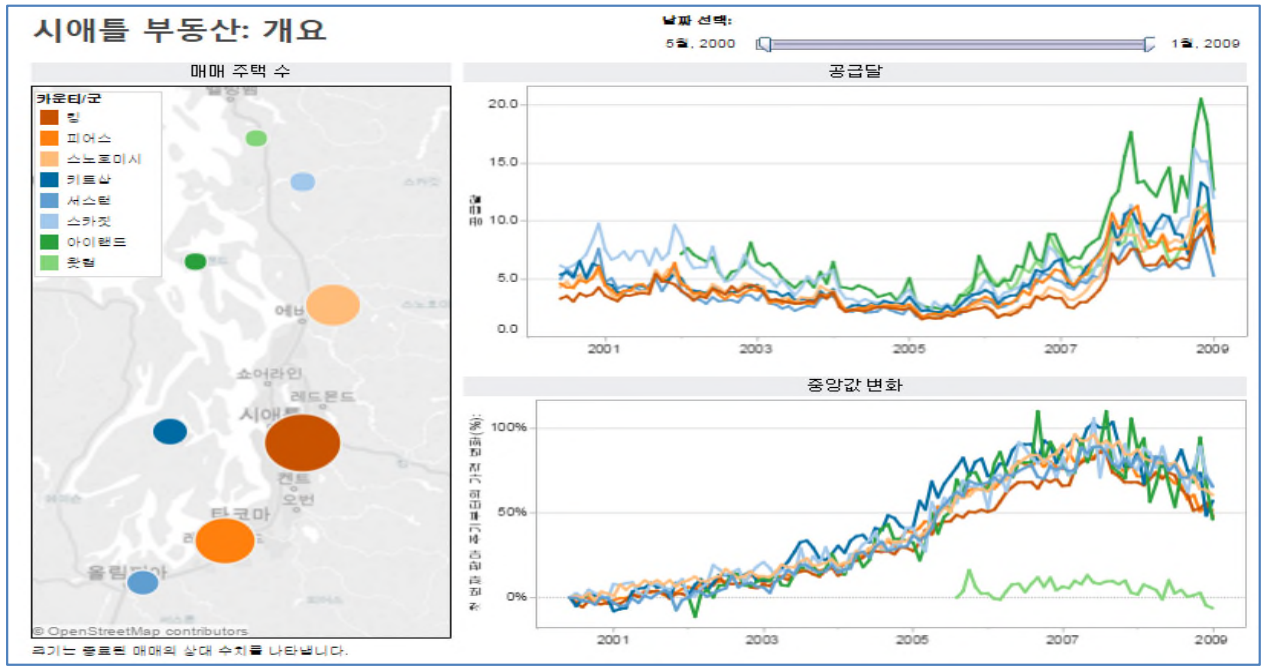
지지도 50% 이상의 순차 패턴은 ?

{겨울연가} => {아폴로13} and

{겨울연가} => {캐스트웨이}

시계열 분석

- **시계열 분석의 사례**
  - 주식의 폐장 가격은 주말마다 발생하는 사건이며, 일정기간 동안의 폐장 가격은 시계열을 형성함
  - 시계열에 대하여 (부분) 시퀀스를 발견함으로써 주가 분석과 예측 가능
- 시애틀 부동산 데이터



회귀분석

- 회귀
  - 회귀는 분류규칙의 특별한 응용으로 많은 연구 분야에서 데이터를 분석하기 위하여 널리 사용되는 일반적인 도구임
  - 분류 규칙이 변수들에 대한 함수이고, 특히 그 변수들을 목표 클래스의 변수로 매핑한다면, 그 분류 규칙을 **회귀규칙(regression rule)**이라고 함
  - 예 :

한 환자에 대하여 n번 연속된 테스트로부터 얻어진 결과값들을 투플로 저장한다고 하자 : (patientID, test_I, test_2,...,test_n)

환자의 생존 확률을 P라고 하고, P=f(test_I, test_2,...,test_n)인 함수 f를 회귀함수하고 부른다. 함수 f가 도메인 변수 test_i에 대하여 선형인 경우에 f를 유도하는 과정을 선형회귀(linear regression)라고 부른다.

신경망 모델

- 신경망 모델
  - 샘플 집합으로부터 커버-핏팅 접근법(curve-fitting approach)을 이용하여 적절한 함수 (선형 및 비선형)를 추론하고, 이를 이용하여 데이터를 분류함
  - **지도 신경망(supervised network)**과 **자율 신경망(unsupervised network)으로 분류할 수 있음**
  - 신경망 모형들은 특정의 문제에 관한 정보로부터 학습을 하는 자체 적응적(self-adapt)임
  - 다양한 문제에 적용할 수 있으며, 데이터의 잡음에 대하여 비교적 견고하고, 다양한 소프트웨어 패키지가 개발되어 있음
  - 결과에 대한 설명이 어려움

유전자 알고리즘

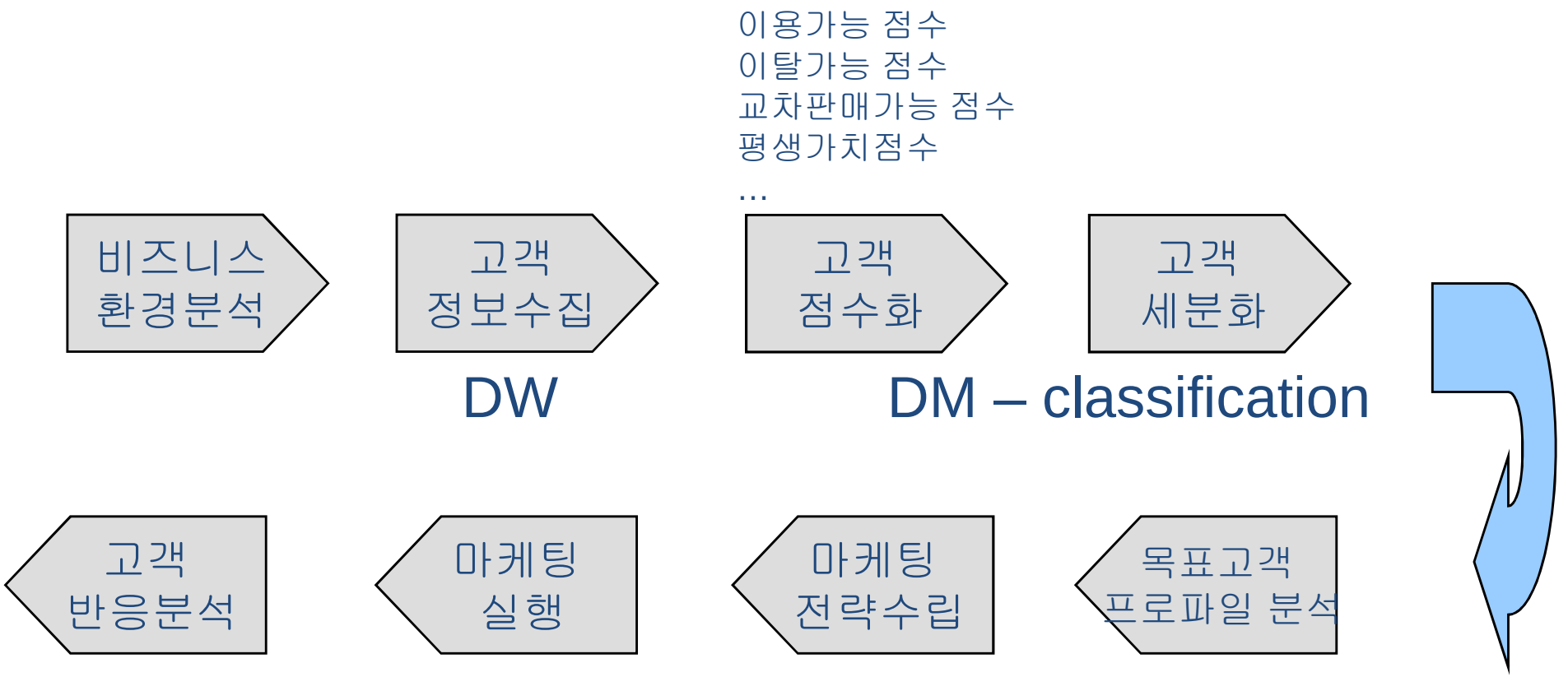
- 유전자 알고리즘(Genetic Algorithm)
  - 탐색 공간이 매우 큰 경우에 좋은 효과를 내는 임의 검색 기법(randomized search procedures)의 한 종류
  - 인간의 유전 개념을 모방한 알고리즘으로 부모개체(문자열)로부터 절단과 병합의 교차연산(cross-over operation)을 수행하면서 새로운 개체(해)를 생성해 나가면서 원하는 정도의 해를 구함
  - 최근에는 데이터 마이닝을 위한 강력한 툴로 활용되기도 함

마이닝 응용

- 데이터 마이닝의 응용 분야
  - 마케팅 - 고객의 구매패턴에 기반을 둔 고객 성향 분석, 광고와 점포의 위치 및 타겟 메일링 등을 포함하는 마케팅 전략의 수립, 고객과 상점 및 상품의 분류, 그리고 카탈로그와 상점 레이아웃 및 광고 캠페인 등의 디자인에 사용된다.
  - 금융 - 고객의 신용가치 분석과, 계좌 분류, 주식이나 채권 및 투자 신탁과 같은 금융 투자 분석, 금융 옵션들의 평가, 사기 행위 적발 등에 사용된다.
  - 제조 - 기계, 인력, 재료와 같은 자원들의 최적화, 제조과정의 최적 설계, 작업 현장 구조 개선, 제품 디자인 (예를들어 고객의 요구사항을 반영한 자동차 설계) 등에 사용된다.
  - 의료 - 치료에 대한 효과 분석, 병원 내에서 치료 과정의 최적화, 환자의 약의 부작용 분석, 유전자를 이용한 제약 개발과 질병치료 등에 이용된다.

마이닝 응용

- 데이터 마이닝의 활용 과정 (CRM의 예)



마이닝 응용

- 대부분의 마이닝 도구들은
  - 연관규칙, 클러스터링 및 분류, 신경망 모형, 연속패턴, 통계적 분석의 기법 등을 지원함
  - ODBC 표준 인터페이스를 이용하여 데이타베이스에 접근하므로 다양한 DBMS와 연계되어 사용 가능함
  - 사용자 인터페이스로는 대부분 정교한 시각화 기술을 가진 GUI가 사용됨
  - 선택적으로 응용 프로그래밍 인터페이스를 제공함 : C 라이브러리와 동적 링크 라이브러리들(dynamic link libraries; DLLs)

마이닝 응용

- 향후 기술동향
  - 현재 데이타 마이닝에서 빠른 처리는 클라이언트-서버 아키텍처, 병렬 데이타베이스, 데이타 웨어하우징 등에서 분산 처리와 같은 최신 데이타베이스 기술을 이용하여 수행되고 있음
  - 데이타 마이닝과 인터넷 기술의 밀접한 결합이 이루어질 것으로 보임
  - 마이닝 도구에서 대규모 데이타 집합들을 다룰 수 있도록 해야 하며, ODBC 표준을 사용하여 다양한 데이터 소스로부터 자료를 수집하여 분석할 수 있어야 함
  - 또한 데이타 마이닝을 위한 소스 데이타로서 이미지와 멀티미디어 데이타 등을 포함시키는 것도 중요한 과제이나 아직 멀티미디어 데이타를 대상으로 하는 데이타 마이닝 기술은 상용화될 만큼 성숙되지는 못한 상황임
  - 최근들어 빅데이터와 마이닝의 결합 => 딥러인, 알파고

마이닝 도구들

회사	제품	기술	플랫폼	인터페이스*
Acknosoft	Kate	결정 트리, 사례 기반 추론	윈도우 NT UNIX	마이크로소프트사의 액세스
Angoss	Knowledge Seeker	결정 트리, 통계학	윈도우 NT	ODBC
Business Objects	Business Miner	신경망, 기계 학습	윈도우 NT	ODBC
CrossZ	QueryObject	통계적 분석 최적화 알고리즘	윈도우 NT MVS UNIX	ODBC
Data Distilleries	Data Surveyor	포괄적 Can Mix 데이터 마이닝	UNIX	ODBC ODMG-호환
DBMiner Technology Inc.	DBMiner	OLAP 분석, 연관, 규칙, 군집화 알고리즘		Microsoft 7.0 OLAP MGr
IBM	Intelligent Miner	분류, 연관 규칙, 예측 모델	UNIX (AIX)	IBM DB2
Megaputer Intelligence	Polyanalyst	기호 지식 획득, 진화적인 프로그래밍	윈도우 NT OS/2	ODBC Oracle DB2
NCR	Management Discovery Tool (MDT)	연관 규칙	윈도우 NT	ODBC
SAS	Enterprise Miner	결정 트리, 연관 규칙, 신경망, 회귀, 군집화	UNIX (Solaris) 윈도우 NT Macintosh	ODBC Oracle AS/400
Silicon Graphics	MineSet	결정 트리, 연관 규칙	UNIX (Irix)	Oracle Sybase Informix