Research on Evolution of data Mining Systems

2003. 9. 4.

Han-joon Kim
University of Seoul

1

Data Mining

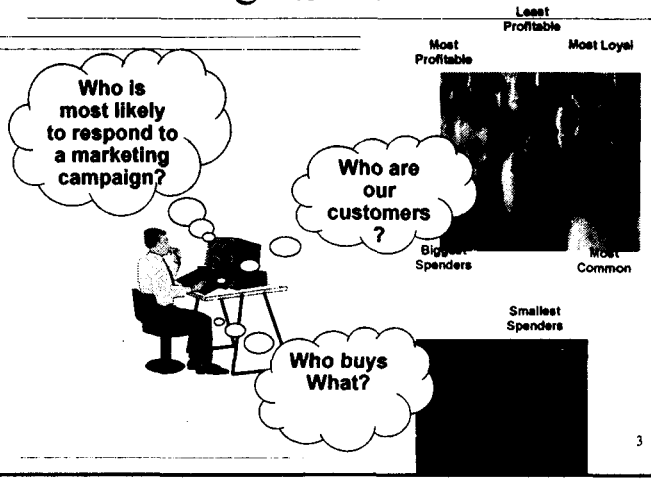
- Uncover the hidden pattern from massive data (data warehouse)

Build a reasonable model to predict the future for business advantage

Decision Making based on the learned models

2

Data Mining: Typical Questions



3

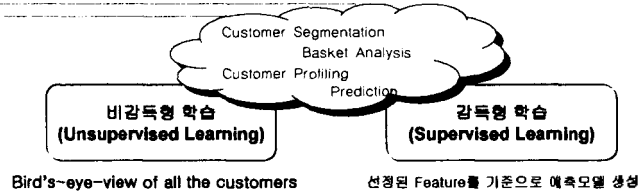
Business Applications of Data Mining

- **Industries**
  - Retail/Marketing, Telecommunications, Insurance, Health-care, etc
- **(e-)CRM**
  - Customer creation/retention/churn prediction
  - Personalization
  - Recommendation
- **Insurance/Banking/Finance**
  - Targeted marking
  - Risk management: fraud detection
- **E-commerce**
  - Automated classification of electronic Information
    - such as customer complaint, product, and e-mail.

4

Data Mining:

Two types of learning method/model



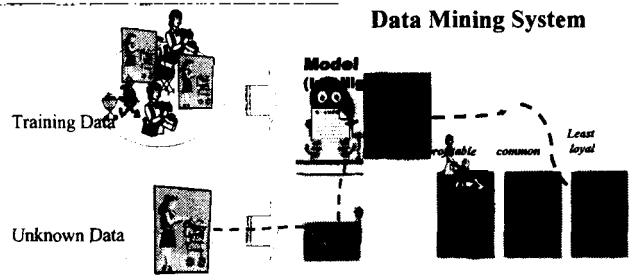
Bird's-eye-view of all the customers

선정된 Feature를 기준으로 예측모델 생성

설명모델
(Descriptive model)

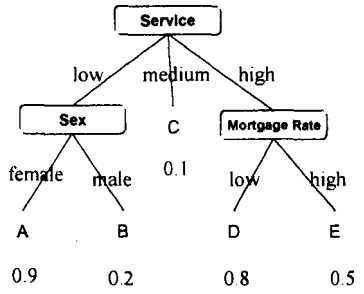
예측모델
(Predictive model)

Data Mining Models



Training Data → Prediction Model

A Business Model of Data Mining



Customer Loyalty Model

- Service Level: low
- Sex: Male
- Rate: Low

Customer prediction:
B: Low loyalty

Customer retention:
B→D

Business Models of Data Mining

- **Segmentation model**
  - **Partition the customer base into mutually exclusive groups**
- **Churn Prediction model**
  - **Identify customers who are at risk of dropping of switching their service to other providers**
- **Acquisition model**
  - **Acquire new customer more cost-effective**
- **Cross/Up-sell model**
  - **Identify customers who are the best prospects for the purchase of additional products**
- **...**

Operational Problems of Data Mining Systems (Tools)

- **Building a (near-to) perfect model is very difficult**
  1. **Big effort in obtaining sufficient training data**
  2. **Dynamically changing environment (Data, Customers, ...)**
     - **Customer purchase behavior, credit card transactional flow**
     - **Network event log, telephone call records**
  3. **Concept Drifts**
     - **Time-evolving trends: ex) User interest**
  4. **Time-critical decision with massive data stream**
  5. **Generation of new categories (customer groups)**

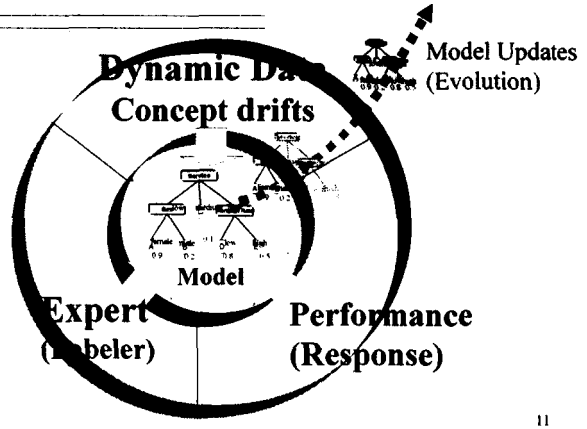
9

Business Models of Data Mining

- **Therefore,**
  - **Require "continuous update of models"**
- **Factors to be considered when building the models**
  - **What features?**
    - **Demographic data, behavior data, etc**
  - **What training data?**
    - **How many?**
    - **Which one is better?**
  - **Decision Making (Action)**
    - **Should be Accurate**
    - **Should be Cost-efficient**

10

Factors that incurs Model Updates (Evolution)



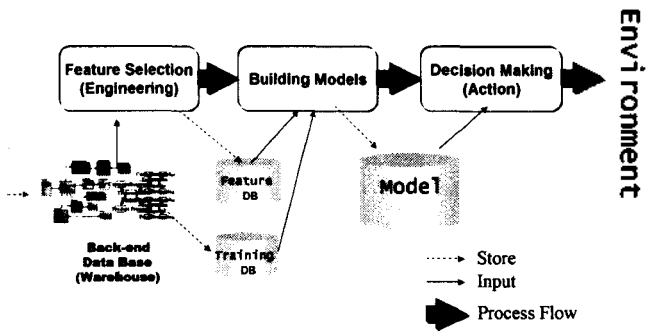
11

Issues on Model Updates

Training Data		Concept Drift		Performance, Response-sensitive
Small data	On-line Env.	Linear	Non-linear	
EM algorithm	EM + AL	Forgetting (Windowing)	Conceptual Clustering	• Cost-sensitive learning • Reinforcement learning
Active Learning (Selective sampling)				
Incremental learning				

12

Data Mining System



13

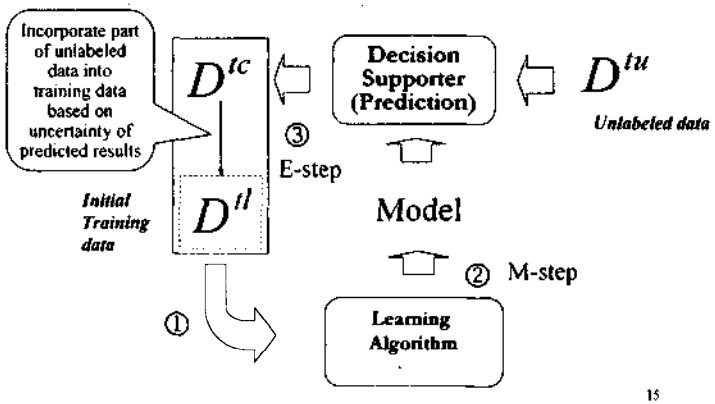
Building models with small training data

- **Using EM algorithm with a small number of labeled data and a large number of unlabeled data**
  - **Improve the conventional EM algorithm with prediction uncertainty measure (*)**
- **Uncertainty**
  - **The degree of uncertainty in the prediction of the unknown example with respect to the current model derived from given training data**

* Han-joon Kim, Jae-young Chang, Improving Naive Bayes Text Classifier with Modified EM Algorithm, 14th International Symposium on Methodologies of Intelligent Systems, 2003 (To appear in LNAI Vol. 2871)

14

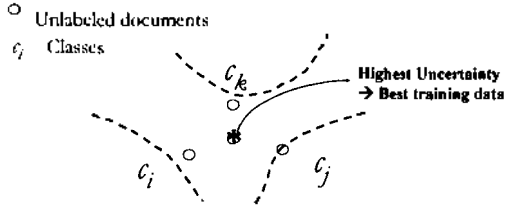
Modified EM algorithm



15

Active Learning

- **What is the best candidate data for learning?**
  - **cf) Random sampling**
  - **Uncertainty-based sampling**
    - **The performance depends on the type of uncertainty measures**
  - **Committee-based Sampling**



16

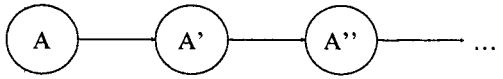
Incremental learning

- **Batch learning**
  - Provide all of training examples for building a model
  - Impractical in real world
  - May give incorrect training examples
- **Incremental learning**
  - Building a model with currently available examples
  - Learning methods for incremental learning
    - Should not destruct the current model
    - - Naïve Bayes algorithm
    - - Support vector machine (SVM) algorithm
    - cf) decision trees

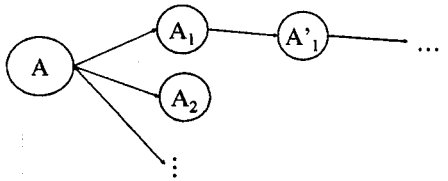
17

Concept drifts

- **Linear Concept drift**



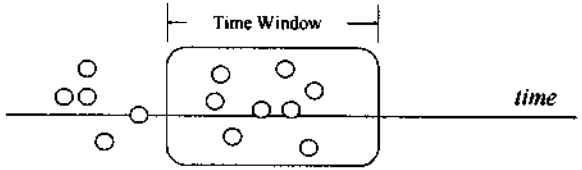
- **Non-linear Concept drift**



18

Monitoring Concept Drifts

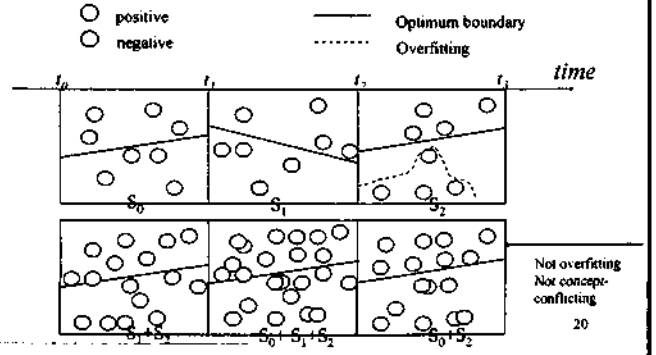
- **Forgetting**
  - **Window-based**
    - Recent information is more trustworthy that older instances
    - Old examples are forgotten (erased from memory)



19

Monitoring Concept Drifts

- **Forgetting**
  - **Considering data distribution**



20

Monitoring Concept Drifts

- **Considering data distribution**
  - **When noise patterns appear**
    - **Then, forgetting should be delayed only if conflicting concepts do not happen.**
  - **When a more general concept is defined**
    - **When a certain specific concept becomes salient, the concept can be isolated (defined)**
    - **Conventionally, use conceptual clustering algorithms to isolate the concept**

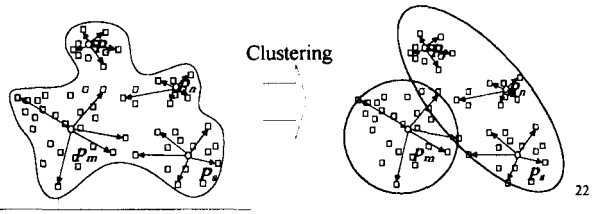
21

Monitoring Concept Drifts of Classes

- **Detecting Concept drift**
  - **Compute the heterogeneity of each class using distance functions**

$$H(c_i) = \sum_{d \in c_i} \frac{cdist(d, c_i)}{|c_i|}, \text{ where } cdist(d, c_i) = \min_{p \in c_i} dist(d, p)$$

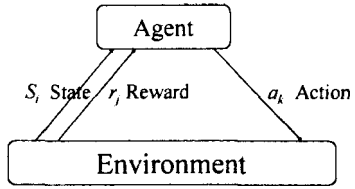
- **Clustering**
  - **Discover new classes (topics)**



22

Reinforcement Learning

- **Goal**
  - **Learn to choose actions that maximize the cumulative rewards ($r_0 + \gamma r_1 + \gamma^2 r_2 + \ldots$, where $0 < \gamma < 1$).**



$$S_0 \xrightarrow[r_0]{a_0} S_1 \xrightarrow[r_1]{a_1} S_2 \xrightarrow[r_2]{a_2} \ldots$$

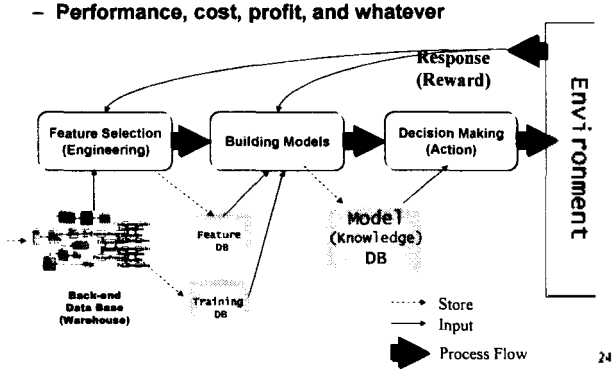
Ex) Targeted Marketing

agent	*Recommender*
Environment	*Customer*
State	*State*
Reward	*Marketing Cost* *Net profit* *Performance*
Action	*Targeted Marketing*

23

Model Updates using Reinforcement Learning

- **Reward as another features**
  - **Performance, cost, profit, and whatever**



24

Conclusions

- Towards unified framework for model updates of Operational Data Mining systems
  - Feature engineering
  - Model building (Learning)
    - Focus on more accurate models
  - Decision supporting
    - Consequently, time/Cost-saving or profitable model should be constructed
- Related research area
  - Active learning
  - Incremental/On-line learning
  - Reinforcement learning
    - Response (reward) → feature of models
  - Model monitoring
  - Evolving (training) data monitoring
    - Data expiration problem

25