Course Outline

- Introduction
- Frequent Patten
- Classification
- Cluster Analysis
- Outlier Detection
- Data Warehouse and OLAP Tech for Data Mining
- Data Mining

Reference Book

Data Mining : Concept and Techniques [Jiawei Han]
Priciples of Data Mining [David J. Hand]
数据仓库与数据分析原理 [王珊]

Concept

> Key Words : Data ,Information, Knowladge [Know The Difference]

`Data Mining`
Extraction of **interesting** patterns or knowledge from huge amount of data

> Objective vs. subjetive intertesingness measures
>
> Objective : based on statistics and structures of patterns
>
> Subjectve : based on user's belief in the data

`KDD Process`
(Data) - 数据集成 — 数据预处理 — 数据挖掘 — 评估表示 - (Knowledge)

Database

- Relational database
- Data warehouse
- Transaction database
- *Object-relational database*
- *Temporal database and time-series database*
- *Text database and multimedia database*
- ...

数据挖掘的特点

1. 真实
2. 海量
3. 随机查询

4. 发现潜在知识

Data Mining Functionalities

- Concept Descripition
- Association
- Classification and Prediction
- Cluster analysis
- Outlier analysis
- Trend and evolution analysis
- Other Pattern Detection

Generalized Framework for Data Mining

- Techniques (本次课程重点)
    - Association rule discovery
    - Sequential pattern discovery
    - Cluster analysis
    - Outlier Detection
    - Classifier Building
    - Data Cube / Data Warehouse Construction
    - Visualization
- Applications［应用到不同的领域］
- Principles［基础能力］
    - Database Technology
    - AI / ML
    - Statistics
    - Information Theory

数据挖掘算法

1. 聚类分析
    - 基于 划分 / 层次 / 密度 / 方格 / 模型 的算法
2. 分类分析
    - 决策树 / 贝叶斯 / SVM / 神经网络

数据挖掘组件化思想

1. 模型（model）或模式（pattern）结构
    - 模型－全局
    - 模式－局部
2. 数据挖掘任务
    - 模式挖掘（项集 / 子序列 / 子结构）
    - 描述建模（eg. Clustering）
    - 预测建模 （eg. Regression / Classification）

3. 评分函数 （似然 / 误差平方和 / 准确率 / 召回率 / F1）
4. 搜索和优化方法（确定模型结构及其参数值）
    - 优化方法（Hill-Climing / Steepest-Descend/Expectation-Maximization）
    - 搜索方法（贪婪 / 分支 / 深度 / 宽度）
5. 数据管理策略

相关链接

Association rule learning [Include FP-Growth]