机器学习与数据挖掘 (研) 课程笔记*

戴唯思

1 Overview

1.1 Course Overview

Qinbao Song, 82668645-731, w1-building Office 731, qbsong@mail.xjtu.edu.cn
数据泛滥: 生成的数据更多, 被捕获的数据更多
问题: Data rich, knowledge poor, 并非所有的数据都是信息, 并非所有信息都是知识
data → useful knowledge
解决方案: 机器学习, 数据挖掘.

1.2 Cource Objects

对机器学习和数据挖掘基本概念和基本原理的比较深入的了解, 从理论上对多种模型和算法的理解, 实现 ML 和 DM 算法的实际动手经验.

涉及 Association, Classification & Prediction, Clustering, Various algorithms for implementing the functionalities

1.3 Distinguishing between ML & DM

1.3.1 Machine Learning

Address the question of how to build computer programs that improve their performance at some task through experience. Learning is a process → algorithm/program 搜索的问题, 需要技巧.

1.3.2 Data Mining

Extract interesting knowledge from huge anoumts of data stored in databases, data warehouse, and other information repositories 从数据仓库里抽取感兴趣的知识. 发现隐藏的事实. 知识: 规则, 规律, 模式, 约束. 是**知识发现**的必要步骤 (Data consolidation, selection & preprocessing, data mining, interpretation & evaluation)

1.3.3 Differences

ML 着重理论, DM 着重实践

*基于课堂笔记整理, 依据课件和更多资料修正和补充. 授课教师: 宋擒豹. 编译日期: September 15, 2011

1.3.4 Stat.

统计更理论, 更重视检测假说; 机器学习 more heuristic(启发式), 在一步的结果基础上推测下一步, 如何提高性能, 超出数据挖掘的领域.

1.4 What can ML & DM do

应用: 目标识别, 文本分类, 语音识别, 传感器数据建模, 自动驾驶, 用户兴趣学习, 疾病诊断,

1.5 Resources

Journal of Machine Learning Research, Machine Learning, Data Mining and Knowledge Discovery, IEEE Transactions on Knowledge and Data Engineering, ACM Transactions on Knowledge Discovery from Data, ..., KDD, ICDM, ICML, ..., UCI Repository, UCI KDD Archive, Statlib, Delve, ...

Software packages: MLC++, GALIB(MIT), Weka(Data Mining in Java)

1.6 Text Book

Introduction to Data Mining(数据挖掘导论), Pang-Ning Tan.
Principles of Data Mining

1.7 Rules & Policies

No cheating tolerated. Late assignments not accepted.

1.8 Grading

- Attendance – 10%
- Project – 30%
  - Completeness – 34%
  - Clarity – 33%
  - Creativity – 33%
- Final Exam – 60%
  on Dec 12, 2011, non-open-book

2 Why data mining

> Necessity is the mother of invention. (需求是发明之母)

解决数据爆炸问题. 数据 → 知识. Data warehousing: 建立数据仓库, 将数据经过一定处理后保存起来, 供 OLAP(on-line analytical processing, 联机分析). **感兴趣**的数据: **valid, novel, useful, understandable** Data mining 的同义词: information harvesting, knowledge mining, data archaeology, knowledge extraction, siftware, data dredging, database mining, data pattern processing, knowledge discovery in database. 对数据挖掘的需求增长, 出现成熟的软件包. 竞争压力.

2.1 Data Mining Development

Associated fields: Statistics, machine learning, database, algorithm, information retrieval (, visualization, other disciplines)

- Information Retrieval: similarity measures, hierarchical clustering, IR systems, imprecise queries, textual data, web SEs
- Statistics: Bayes theorem, regression analysis, EM algorithm K-mean slustering, time-series analysis
- ML: neutral networks, decision tree algorithms
- Algorithms: algorithm design techniques, algorithm analysis, DS
- DB: relational data model, SQL, assiciation rule algorithms, data warehousing, scalability techniques

Traditional techniques may be unsuitable due to ...

- Enormity of data
- High dimensionality of data
- Heterogeneous, distributed nature of data

2.2 On what kinds of data

- Traditional DB and Applications
  relational db, data warehouse, transactional db
- Advanced db and advanced apps
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data (incl. biosequences)
  - Structure data, graphs, social networks and link DBs
  - Object-relational DBs
  - Heterogeneous DBs and legacy DBs
  - Spatial data and spatiotemporal data
  - Multimedia database
  - Text DBs
  - WWW

2.3 Classification Schemes

Classifying by general *functionality*:

- **Descriptive** data mining — patterns
  clustering, summarization, association rules, sequence discovery
- **Predictive** data mining — value
  classification, regression, time-series analysis, prediction

By different *views*:

- kinds of data to be mined
  relational, data warehouse, transactional, stream, object-oriented/relational, active, spatial, ...
- kinds of knowledge to be discovered
  characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis...
  multiple/integrated functions and mining at multiple levels
- kinds of techniques utilized
  DB-orienteed, data warehouse (OLAP), ML, statistics, visualization, ...
- kinds of applications adapted
  retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, ...

2.4 Potential Applications

- Data analysis and decision support
  Market analysis and management, risk ..., fraud detection
- Other
  Text mining, stream data mining, bioinformatics and bio-data analysis

3 KDD

KDD — **Knowledge Discovery in Data**: selection and processing of data for automated discovery of *novel, accurate and useful* patterns and the modeling of real-world phenomena.

3.1 The Virtuous Cycle

1. Identify problem or opportunity — **Problem**
2. The KDD process — **Knowledge**
3. Act on knowledge — **Results**
4. Measure effect of action — **Strategy**
5. *(Cycling .....)*

3.2 Roles in the KDD Process

- Domain experts 领域专家
- Mining specialists 数挖人才
- Data Admin 数据管理员

Process: business analysis → data analysis → data gathering → data preparation → data mining → result interpretation → business application → business feedback → BUSINESS GOAL

4 Steps in KDD: → (data sources) → Data consolidation → (consolidated data) → selection & preprocessing → (prepared data) **data mining** → (patterns and models) → interpretation & evaluation → (knowledge)

1. Learning the application domain
   relevant prior knowledge and goals of application
2. Data consolidation: creating a target data set
3. Selection and preprocessing
   data cleaning (may take 60% of effort), data reduction and projection
4. Choosing functions of data mining
   summarization, classification, regression, association, clustering
5. Choosing the mining algorithm(s)
6. **Data mining**: search for patterns of interest
7. **Interpretation and evaluation**: analysis of results
   visualization, transformation, removing redundant patterns, ...
8. Use of discovered knowledge

3.2.1 Core Problems & Approaches

Problems:

- **Identification** of relevant data
- **Representation** of data
- **Search** for valid patterns or models

Approaches:

- Top-down **deduction** by expert
- Interactive **visualization** of data/models (OLAP)
- *Bottom-up induction from data*

3.3 Data Consolidation and Preparation

Garbage in → garbage out, 数据质量很重要

The quality of results relates directly to quality of data. 50%-70% of KDD process effort is spent on this step.

3.3.1 Data Consolidation

From internal and external data sources to consolidated data repos (incl. object/relation DBMS, multidimensional DBMS, deductive database, flat files)

1. Determine preliminary list of attributes
2. Consolidate data into working DB
3. Eliminate or estimate missing values
4. Remove *outliers* (obvious exceptions)
5. Determine prior probabilities of categories and deal with *volume bias*

3.4 Data selection and preprocessing

Tasks:

- Generate a set of examples
  choose sampling method, consider sample complexity, deal with volume bias issues
- Reduce attribute dimensionality
  remove redundant and/or correlating attributes, combine attributes (sum, multiply, difference)
- Reduce attribute value ranges
  group symbolic discrete values (example: generations of program languages), quantize continuous numeric values
- Transform data de-correlate and normalize values (examples: units), map time-series data to static representation

OLAP and visualization tools play key role

3.5 Data Mining Techniques

Classfication, cluster analysis, instance based learning, frequent pattern & association rule mining, graph mining, ensemble learning

3.5.1 Classfication: 2-steps

1. Model construction: describe a set of *predetermined* classes
   Training dataset: tuples for model construction
   Classification rules, decision trees, math formula

2. Model application: classify *unseen* objects
   Estimate accuracy of the model using an independent test set
   Acceptable accuracy → apply the model to classify tuples ...

Supervised vs Unsupervised learning:

- Supervised learning (classification)
  Supervision: objects in the traning data set have labels
  New data is classified based on the training set

- Unsupervised learning (clustering)
  The class labels of training data are unknown

Cluster: a collection of data objects, similar to one another within the same cluster, dissimilar to the objects in other clusters

Cluster analysis: finding clustering such that the objects in a cluster will be similar to one another and different from the objects in other clusters based only on information found in the data that describes the objects and their relationships. Intra-cluster distances → minimized

Cluster Analysis Techniques:

- **Partitioning** algorithms: partition the objects into k clusters, iteratively reallocate objects to improve the clustering

- **Hierarchy** algorithms: agglomerative 汇聚式, divisive 合并式

- Density-based methods, grid-based methods, ...

Instance based learning: lazy, processing is postponed, until queries are encountered, stores all training examples. Forms: k-NN (k-nearest neighbor, choose k of the "nearest records"), locally weighted regression, radial basis function, case-based reasoning (符号推理, 机械设计中常用, 易与 k-NN 混淆)