SIT409: Data Mining and Knowledge Discovery

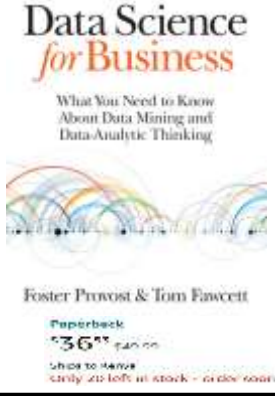
SCO415: Data Warehouse and Data Mining

A Karwega

Course Topics

01. Data Mining Fundamentals
02. Data Mining: Learning Methods
03. Data warehousing and OLAP
04. Data Mining Algorithms
05. Data Mining Process
06. Review and Selection of Commercial Tools for Data Mining

Key Reference Texts



01. Data Mining Fundamentals

- What is Data Mining?
- Tasks Achievable through Data Mining;
- Motivations behind Data Mining;
- Case Study

Background: Quotes from business

- "We are sinking in data and starved of information"
- "We are helping clients harness their vast amounts of customer and operational data"
- "We cannot find enough new graduates with the right quantitative skills"
- "We compete on the basis of better knowledge of our customers, using analytics"
- "The riskier our business problems the more we rely on analytics"

Background: Quotes from business

- "After implementing our ERP system we are mining that data, and using data better in different ways"
- "Do you think that, or do you know that?"
- "Those who succeed with six sigma, and then advance in our company, have the better quantitative skills"
- "We are basing our strategy on analytics, especially customer analytics"

Data Mining Defined

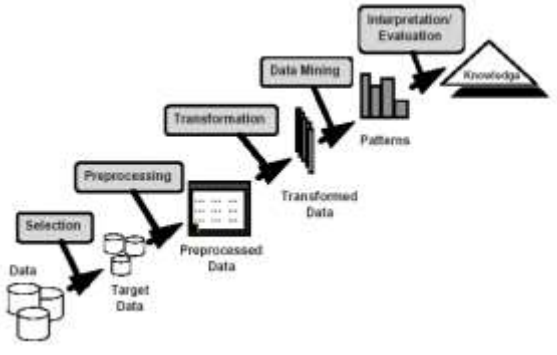
- a.k.a knowledge discovery in databases (KDD)
- "the process of exploration and analysis, by automatic or semi-automatic means of large quantities of data in order to discover meaningful patterns and results"

(Berry and Linoff, 1997,2000)

- "The non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data"

(Fayad, shapiro et al)

The DM Process – Common Stages



What is (not) Data Mining?

- What is not Data Mining?
  - Look up phone number in phone directory
  - Query a Web search engine for information about "Amazon"
- What is Data Mining?
  - Certain names are more prevalent in certain Kenyan locations (Omondi, Otieno, Onyango… Nyanza)
  - Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com,)
    - Note that though some part of the name appears similar, These are completely different contexts

Some DM Techniques

Contributing Disciplines

Why Mine Data?

- **To Uncover Hidden Relationships**
- **Find something unusual or unexpected**
- **Improve upon domain expert's knowledge**
- **Manage large data sets**
- **Create Predictive Analytics platform**
- **Maximize value of Data**
- **Competitive Edge**

Data Mining: Why Now?

- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - Purchases at department/ grocery stores
  - Bank/Credit Card transactions
- Computing power has become *very much* more available and affordable.
- Competitive Pressure is Strong
  - Need to provide better, customized services for an *edge* (e.g. in Customer Relationship Management)

Data Mining: Why Now? (Scientific viewpoint)

- Data collected and stored at enormous speeds (GB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data
- Traditional techniques infeasible to process raw data:
  - Enormity of data
  - High dimensionality of data
  - Heterogeneous, distributed nature of data

Data Mining Tasks

1. **Classification: learning a function that maps an item into one of a set of predefined classes**
2. **Regression: learning a function that maps an item to a real value**
3. **Clustering: identify a set of groups of similar items**
4. **Dependencies and associations: identify significant dependencies between data attributes**
5. **Summarization: find a compact description of the dataset or a subset of the dataset**

Data Mining Methods

1. **Decision Tree Classifiers**: Used for modeling, classification
2. **Association Rules:** Used to find associations between sets of attributes
3. **Sequential patterns**: Used to find temporal associations in time series
4. **Hierarchical clustering:** used to group customers, web users, etc

Data Pre-Processing: Why?

- Data in the real world is dirty
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - noisy: containing errors or outliers
  - inconsistent: containing discrepancies in codes or names
- Low quality data => Low quality mining results!
  - Quality decisions must be based on quality data
  - Data warehouse needs consistent integration of quality data

Common Reasons for Low Quality Data

- Attributes of interest are not available (e.g., customer information for sales transaction data)
- Data were not considered important at the time of transactions, so they were not recorded!
- Data not recorded because of misunderstanding or malfunctions
- Data may have been recorded and later deleted!
- Missing/unknown values for some data

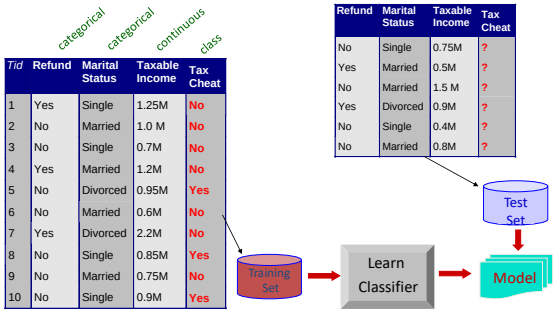
Data cleaning tasks

- Fill in missing values
- Identify outliers and smooth out noisy data
- Correct inconsistent data

Classification: Definition

- Given a collection of records (training set)
  - Each record contains a set of attributes, one of the attributes is the class.
- Find a model for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A test set is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Classification Example



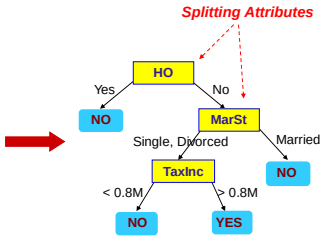
Variable Types

- *Numerical*: Domain is ordered and can be represented on the real line (e.g., age, income)
- *Nominal* or *categorical*: Domain is a finite set without any natural ordering (e.g., occupation, marital status, race)
- *Ordinal*: Domain is ordered, but absolute differences between values is unknown (e.g., preference scale, severity of an injury)

Example of a Decision Tree

Tid	Refund	Marital Status	Taxable Income	Tax Cheat
1	Yes	Single	1.25M	No
2	No	Married	1.0 M	No
3	No	Single	0.7M	No
4	Yes	Married	1.2M	No
5	No	Divorced	0.95M	Yes
6	No	Married	0.6M	No
7	Yes	Divorced	2.2M	No
8	No	Single	0.85M	Yes
9	No	Married	0.75M	No
10	No	Single	0.9M	Yes

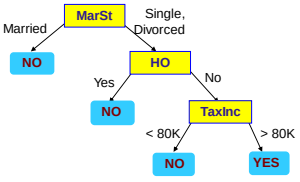
Training Data



Model: Decision Tree

Another Example of Decision Tree

Tid	Refund	Marital Status	Taxable Income	Tax Cheat
1	Yes	Single	1.25M	No
2	No	Married	1.0 M	No
3	No	Single	0.7M	No
4	Yes	Married	1.2M	No
5	No	Divorced	0.95M	Yes
6	No	Married	0.6M	No
7	Yes	Divorced	2.2M	No
8	No	Single	0.85M	Yes
9	No	Married	0.75M	No
10	No	Single	0.9M	Yes



There could be more than one tree that fits the same data!

Other Approaches to Classification

- Linear Discriminant Analysis
- *k*-nearest neighbor methods
- Logistic regression
- Neural networks
- Support Vector Machines

Classification: Application 1

- Direct Marketing
  - Goal: Reduce cost of mailing by targeting a set of consumers likely to buy a new cell-phone product.
  - Approach:
    - Use the data for a similar product introduced before.
    - We know which customers decided to buy and which decided otherwise. This {buy, don't buy} decision forms the class attribute.
    - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
      - Type of business, where they stay, how much they earn, etc.
    - Use this information as input attributes to learn a classifier model.

Classification: Application 2

- Fraud Detection
  - Goal: Predict fraudulent cases in credit card transactions.
  - Approach:
    - Use credit card transactions and the information on its account-holder as attributes.
      - When does a customer buy, what does he buy, how often he pays on time, etc
    - Label past transactions as fraud or fair transactions. This forms the class attribute.
    - Learn a model for the class of the transactions.
    - Use this model to detect fraud by observing credit card transactions on an account.

Classification: Application 3

- Customer Attrition/Churn:
  - Goal: To predict whether a customer is likely to be lost to a competitor.
  - Approach:
    - Use detailed record of transactions with each of the past and present customers, to find attributes.
      - How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
    - Label the customers as loyal or disloyal.
    - Find a model for loyalty.

Clustering Definition

- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - Euclidean Distance if attributes are continuous.
  - Other Problem-specific Measures.

Clustering

- Output: (k) groups of records called clusters, such that the records within a group are more similar to records in other groups
  - Representative points for each cluster
  - Labeling of each record with each cluster number
  - Other description of each cluster
- An unsupervised learning process: No record labels are given to learn from
- Usage:
  - Exploratory data mining
  - Preprocessing step (e.g., outlier detection)

Clustering Application: Market Segmentation

- Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
- Approach:
  - Collect different attributes of customers based on their geographical and lifestyle related information.
  - Find clusters of similar customers.
  - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Association Rule Discovery

- Given a set of records each of which contain some number of items from a given collection;
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:
{Milk} --> {Coke}
{Diaper, Milk} --> {Beer}

Challenges of Data Mining

- Complex and Heterogeneous Data
- Scalability, Dimensionality
- Data Quality, Data Formats
- Data Ownership and Distribution
- Ethics, Privacy Preservation, Political Correctness
- Streaming Data
- …