Data Mining
Association Analysis: Basic Concepts and Algorithms

Lecture Notes for COMP7650

Review

Review

- **Data**
  - Types
  - Quality
  - Preprocessing
- **Classification**
  - Decision trees
  - Gaussian Mixture
  - Hidden Markov model
  - K-nearest neighbor
  - LVQ
  - MLP, RMLP
- **Clustering**
  - K-means
  - SOM, SOM-SD
- **Projection, Visualisation**
- **Association rules**

Data

- Enormous amounts of data is collected in many areas.
  - How to process large amounts of data (Data Mining)
  - How to detect value added information from data (Knowledge Discovery)
- Types of data can include: Vectors and matrices, sequences, trees and graphs.
- Important properties: Dimensionality, sparsity, relatedness, balance.
- Quality is affected by noise, outliers, and duplicate data
- Techniques to Data Mining and Knowledge Discovery are drawn from Statistics, and Machine Learning (also from DataBase systems).

Classification

- Aim is to map a set of attributes to a class.
- Commonly achieved by using machine learning approaches:
  - Build a model over a given training set, then.
  - Classify unseen data (from a test set)
- Main issues:
  - Overfitting
  - Training parameters
  - Selection of training algorithm

Classification Algorithms

- Gaussian Mixtures (actually a clustering algorithm)
- Bayes Classifier
- Hidden Markov model
- K-nearest Neighbor
- LVQ
- Multi-layer Perceptron Networks
- Recurrent MLP
- Recursive MLP

Which one has problems with unbalanced data?

Which one suffers from long term dependency?

Clustering

- Group data together into the same group if they share a common property or are similar to each other, otherwise group them into different groups.
- Clustering is commonly done when class labels are not available, or when dimension reduction or visualisation is required.
- Main problems: definition of “similarity”, validation of clustering results, number of clusters, overlapping clusters.
- Algorithms:
  - K-means
  - Self-Organizing Maps
  - SOM for Structured Data
  - Hierarchichal Clustering

Projection and Visualization

- Reduce the dimensionality of the input space to a more manageable level while keeping loss of information at a minimum.
- Reduce dimensionality to two (or three) dimensions to allow the visualization.
  - Best is to use topology preserving mapping

Association rules

- Extract rules from an analysis of transaction records.
- Important measures:
  - Support and Confidence
  - Interest measure
- Main problem: Computational complexity!
- Solution: Pruning as much as possible. Note that this does not affect the computational complexity, but rather it keeps the search space small (within a manageable level).

Main messages

- There is no plug-an-play solution to data mining, and knowledge discovery.
- You need to know which approach is most suitable to a given problem.
  - You need to know the strengths and weaknesses of the approaches.
- You need to know how to use the approach (which parameters to use, etc.)

Study hints for the exam

- Know (and understand) the algorithms, their abilities, and uses!
- Know und understand the assignment tasks.
- Exercise the solving of logic functions using MLP.

- There are no programming tasks.

- Exam date and location:

  10/December/2009 at 19:00 – 22:00 in Room FSC501