Suggested Course Projects

Overview

1. Microeconomic Clustering
2. Clustering Microarray Data
3. Subspace Clustering
4. Interpretation of SVM
5. Evaluation of Classification Methods
6. Shape-Based Classification of Spatial Objects

1. Microeconomic Clustering

- Assumption
  - set of possible decisions D of an enterprise
  - set of customers C
  - contribution of customer i to the utility of the enterprise under decision x can be estimated using the available data y_i on customer i: $g(x, y_i)$
  - the enterprise is looking for the decision maximizing the overall utility:

$$\max_{x \in D} \sum_{i \in C} g(x, y_i)$$

- Idea
  - Cluster the customers (e.g., into k clusters)
  - For each cluster, an optimum decision will be taken
- Naive Algorithm
  - Enumerate all possible combinations of k decisions
  - Assign each customer to the decision with the highest utility

1. Microeconomic Clustering

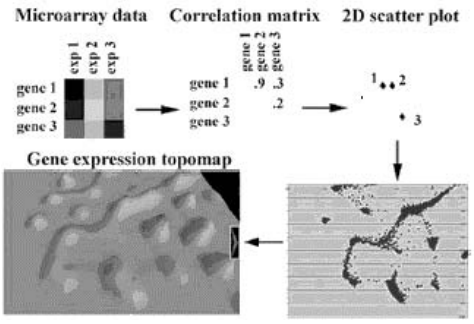
- Challenges
  - Interesting instances of the microeconomic clustering problem are NP-hard
  - Problem formulation does not easily generalize to non-commercial applications
- Possible Projects
  - Experimental evaluation of algorithms from the literature
  - Efficient algorithm for special instances
    e.g., for the catalog segmentation problem
    (form catalogs with a limited number of products)
  - Formulate other clustering problems as a problem of microeconomic clustering
    e.g. text clustering or clustering of Microarray data
    decision = choice of cluster description?

2. Clustering Microarray Data

- Genomic sequence data available, but function of genes unknown
- Microarray data:
  Expression levels for different genes (rows) in different experiments (columns)
- Assumption
  Genes that show very cohesive levels of coexpression in diverse microarray experiments are likely to function together
- Goal: find clusters of genes showing a high level of coexpression
- Two approaches of analysis
  - Query-based clustering (input one set of genes, find descriptive experiments, find similar genes)
  - Classical clustering (unsupervised, find all clusters of genes with similar expression patterns)

2. Clustering Microarray Data

• Existing methods



• Challenges
  – The clustering problem is ill-defined (clarify the expected result)
  – Clusters may exist only in subspaces (using a subset of all experiments)
  – Genes may belong to multiple clusters

2. Clustering Microarray Data

• Possible Projects
  – Experimental comparison of state-of-the-art methods on a microarray dataset
  – Clarify the clustering problem and design a corresponding novel method
  – Modify existing methods for subspace clustering and / or multiple cluster memberships (with / without experimental evaluation)

→ For both, Query-based clustering and Classical clustering

3. Subspace Clustering

- Motivation
  - Clusters may exist only in subspaces
- Existing methods
  - CLIQUE
  - Projected Clustering
  - Fascicles
  - . . .
- Challenges
  - Existing methods need a lot of user parameters (hard to determine)
  - Existing methods can hardly use existing domain knowledge
  - Existing methods are not efficient

3. Subspace Clustering

- Possible Projects
  - Experimental comparison of some state-of-the-art methods on real life dataset
  - Design a novel subspace clustering method based on a different paradigm

    e.g., based on the EM paradigm
  - Design a query-based subspace clustering method

4. Interpretation of SVM

- SVMs achieve very high classification accuracy
- But resulting model is hard to interpret
  Large number of features with relatively high weights
- Goal 1: Explain the classification of an *individual query* object
  Based on a relatively small subset of the set of all features
- Approach 1:
  Consider the query object and „neighboring“ support vectors from both classes
  Identify the features „locally“ distinguishing between the two classes
  Generate an understandable explanation
- Project 1
  Design, implement and experimentally evaluate such a method

4. Interpretation of SVM

- Goal 2:
  Explain the *global characteristics* of both classes
  Approximate the separating hyperplane by a set of hyperrectangles
- Approach 2:
  Perform clustering of both classes using the support vectors as „constraints“
  Approximate the clusters using their bounding boxes
  Generate a description of the two classes based on the bounding boxes
- Project 2
  Design a clustering method that respects such constraints
  Implement and evaluate on some interesting data set

5. Evaluation of Classification Methods

- There are so many different classification methods and many alternatives for data preprocessing . . .
- Classification accuracies / model sizes etc. may vary greatly
- Experimentation with alternative methods requires
  a lot of time
  a lot expertise to choose methods and to evaluate results
- Long term vision:
  Automatically choose model category / model structure and optimize the parameters for classification tasks

5. Evaluation of Classification Methods

- One step towards this vision: develop criteria for the following questions
  (1) For a given classification problem, which feature selection method(s) should be chosen?
  (2) For a given classification problem, which classification methods perform well?
  (3) What classification accuracies / model sizes etc. can one expect?
  (4) What are the dependencies between feature selection and classifier construction?
- Classification problem described by
  – Data characteristics such as number of records, number of attributes, type of attributes, distribution of attributes, class distribution, . . .
  – Classifier quality measures (classification accuracy, precision, recall, . . . and their priorities)

5. Evaluation of Classification Methods

- Approaches
  - Literature study
  - Own experimental evaluation
- Data sets
  - UCI KDD archive http://kdd.ics.uci.edu/
  - KDD Cup datasets http://www.kdnuggets.com/datasets/kddcup.html
  - Other datasets http://www.kdnuggets.com/datasets/index.html
- Outcome of project
  (a) Method for specifying classification problems
  (b) Answers to questions (1) to (4)
    - Select subset of alternative methods to be considered
  (c) Sketch of a method
    - Explore the search space defined by the selected methods
    - How to combine feature selection and classifier construction?
    - How to prune the search space (using expected classifier qualities)?

6. Shape-Based Classification of Spatial Objects

- Given
  Labeled training data set of spatial objects with 2D shapes
  (houses, farms, schools, . . .)
- Wanted
  Classifier that can be predict the class label based on the 2D shape
- Applications
  Interpretation of aerial images
  Generation of prototypical shapes for different object classes
- Standard approach
  Extract features from shapes and apply standard classifiers
- Spatial approach
  Work directly on the shapes
  Develop spatial classifier

6. Shape-Based Classification of Spatial Objects

- Idea for spatial approach
  - Based on version space approach
  - Define version space
  - Define set of generalization operators for shapes
  - Modify existing algorithm (Find-S / Candidate elimination)
- Project
  - Design spatial classifier
  - Implement
  - Evaluate on real data sets (e.g., from Gemure project)