Feature selection for clustering

Victor Kitov

Feature selection for clustering

Problem statement

Select subset of features in which training objects break into distinct clusters in most explicit way.

- Its data mining, not machine learning with exact criterion optimization.
- Categorization of feature selection methods:
    - **Filter methods**: do not rely on particular clustering algorithm
        - generally faster
        - more universal
        - fit less well with exact method
    - **Wrapper methods**: tied to particular clustering algorithm
        - work better for given algorithm

Table of Contents

Features and objects similarity

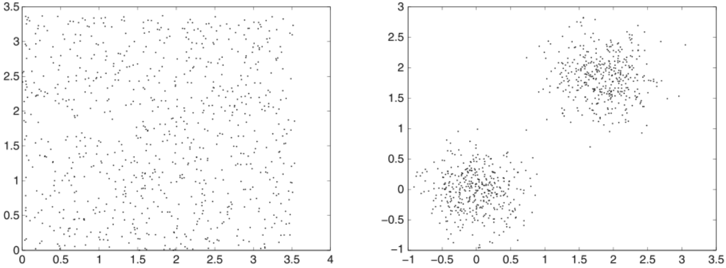
- Intuition: features good for clustering can individually predict well the similarity of objects.
- For 2 randomly chose objects x, x' they should be similar $<=>$ x^i, x'^i are similar.
    - need to define similarity
- Example: news clustering, features-indicators of words:
    - president (indicative for politics cluster)
    - competition (indicative for sports cluster)
    - exhibition (indicative for arts cluster)

Features and objects similarity

- x^i-real feature :
    - $corr(\rho(x, x'), |x_i - x_i'|)$
    - $corr(\mathbb{I}[x \text{ and } x' \text{ are not similar}], |x_i - x_i'|)$
- x^i-binary feature:
    - $corr(\rho(x, x'), \mathbb{I}[x_i = x_i'])$
    - $corr(\mathbb{I}[x \text{ and } x' \text{ are not similar}], \mathbb{I}[x_i = x_i'])$
    - $p(x_i' = 1 | x_i = 1)$ for any x' similar to x.
- Comment: features should have equal scale.

Predictive attribute dependence

- for good clustering feature should be predicted well with other features:

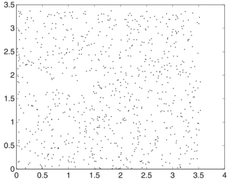


(a) Uniform Data (b) Clustered data

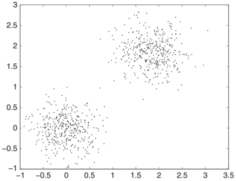
- score of feature i: accuracy of predicting x^i using $\left\{x^j\right\}_{j\neq i}$
- K-NN prediction is preferred due to its geometric intuition

Pairwise distance distribution

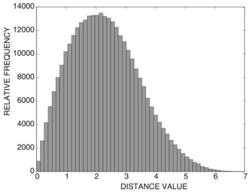
- Estimate distribution of $\rho(x, x')$ for random x, x'
- Good clustering should give multimodal distribution



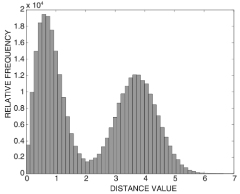
(a) Uniform Data



(b) Clustered data



(c) Distance distribution (uniform)



(d) Distance distribution (clustered)

Pairwise distance distribution

- Consider object representation with features I:$F_I(x) = \left\{x^i\right\}_{i \in I}$
- Possible quality of feature subset I: $Entropy[\rho(F_I(x),\ F_I(x'))]$ for random x, x'.
- Feature subset selection - using backwards suboptimal search:
  - start from full set of features
  - recurrently remove least significant feature, according to $\Delta Entropy$.

Hopkins statistic

Define:

- T - training dataset ($T = \{x_1, ...x_N\}$)
- R - set of real objects $x_1', ...x_K'$
  - each object is selected randomly from T
  - $\alpha_i := \rho(\tilde{x}_i, T)$
- S - set of synthetic objects $\tilde{x}_1, ...\tilde{x}_K$
  - each feature generated randomly independently of others in its domain
  - define $\beta_i := \rho(\tilde{x}_i, T)$
- Hopkins statistic

$$H = \frac{\sum_{i=1}^{K} \beta_i}{\sum_{i=1}^{K} \alpha_i + \beta_i}$$

- $H \in [0.5, 1]$, higher valuer are better

Table of Contents

Wrapper methods

- Filter methods, considered before, do not consider what clustering method will be used
- Wrapper methods do feature selection for particular choice of clustering method.
- Approaches:
  - feature selection with backward search
  - classifier feature selection
- Comments:
  - wrapper methods are tied to final clustering algorithm
  - but filter methods are faster, than wrapper
  - we can use filtering methods to generate candidate feature subsets for wrapper methods.
    - better efficiency

Feature selection with backward search

- Select some cluster evaluation criterion $J(\cdot)$
- Algorithm:

Init $F=\{f^1,...f^D\}$ to contain all features

WHILE clustering quality J(F')-J(F) continues to improve:
    F=F'
    $f'=\arg\max_f J(F\backslash\{f\})$
    set $F'=F\backslash\{f'\}$

RETURN F

Classifier method

- Classifier method:
  1. Perform clustering on $x_1, ...x_N$, obtain cluster labels $c_1, ...c_N$
  2. Use any *supervised* feature selection for $(x_1, c_1), ...(x_N, c_N)$.
- Modifications:
  - apply classifier method iteratively
  - not discard removed features but decrease their weight