Cluster Validity

Erin Wirch & Wenbo Wang

Oct. 28, 2010

Outline

Agenda

- Hypothesis Testing
  - Review of Hypothesis Testing
  - Random Position Hypothesis
  - Random Graph Hypothesis
  - Random Label Hypothesis
- Relative Criteria

Review of Hypothesis Testing

- Test a parameter against a specific value
- Begin with H_0 and H_1 as the null and alternative hypotheses
- Power function:

$$W(\theta) = P(q \epsilon \overline{D}_p | \theta \epsilon \Theta_1)$$

- Goal: make correct decision

Hypothesis Testing in Cluster Validity

- Test whether the data of X possess a random structure
- First step: generate data to model a random structure
- Second step: define a statistic and compare results from our data set and a reference set
- Three methods exist to generate the population under the randomness hypothesis
- Choose best method for the situation

Random Position Hypothesis

- Suitable for ratio data
- Requirement: "All the arrangements of N vectors in a specific region of the l-dimensional space are equally likely to occur."
- This can be accomplished with random insertion of points in the region according to uniform distribution
- Can be used with internal or external criteria

External Criteria

- Impose clustering algorithm on X a priori based on intuitions
- Evaluate resulting clustering structure in terms of independently drawn structure

Internal Criteria

- Evaluate clustering structure in terms of vectors in X
- Example: proximity matrix

Random Graph Hypothesis

- Suitable when only internal information is available
- Definition: NxN matrix A as symmetric matrix with zero diagonal elements
- A(i,j) only gives information about dissimilarity between x_i and x_j
- Thus comparing dissimilarities is meaningless

Random Graph Hypothesis, cont'd

- Let A_i be an NxN rank order without ties
- Reference population consists of matrices A_i generated by randomly iserted integers in the range $[1, \frac{N(N-1)}{2}]$
- H_0 rejected if q is too large or too small

Random Label Hypothesis

- Consider all possible partitions, P' of x in m groups
- Assume that all possible mappings are equally likely
- Statistic q can be defined to measure degree information in X matches specific partition
- Use q to test degree of match between P and P against q_i's corresponding to random partitions
- H_0 rejected if q is too large or too small

Methodology

- To choose the best parameters A for a specific clustering algorithm to best fit the data set X
- Parameter set A
  - the cluster size estimation m
  - the initial estimates of parameter vectors related with each cluster

Method I

- cluster size m is not pre-determined in the algorithm
- criteria: the clustering structure is captured by a wide range of A

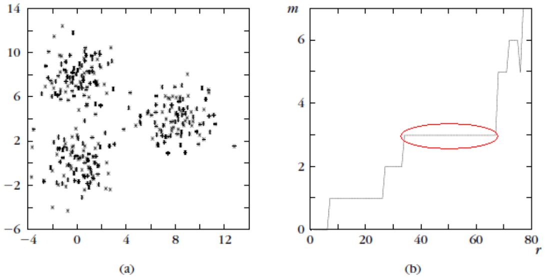

Figure: (a) 2-D clusters from normal distributions with mean $[0,0]^T$, $[8,4]^T$ and $[8,0]^T$, covariance matrices $1.5I$. (b) clustering result (cluster size m) with binary morphology algorithm, with respect of different resolution parameters r

Erin Wirch &
Wenbo Wang

Method I (Cont')

- Comparing by data set with a wider range of variance:

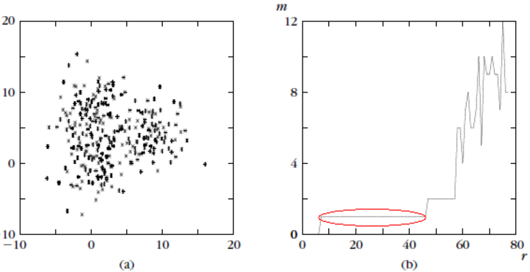

Figure: (a) 2-D clusters from normal distributions with mean $[0, 0]^T$, $[8, 4]^T$ and $[8, 0]^T$, covariance matrices $2.5I$. (b) clustering result (cluster size m) with binary morphology algorithm, with respect of different resolution parameters r

Method II

- cluster size m is pre-determined in the algorithm
- criteria: to choose the best clustering index q in the range of $[m_{min}, m_{max}]$
  - if q shows no trends with respect of m, vary parameter A for each m, choose the best A
  - if q shows trends with respect of m, choose m where significant local change of q happens

Method II (cont')

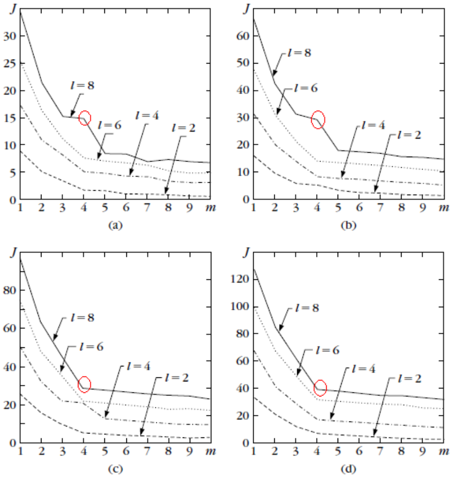

Figure: data set generated from 4 well-separated normal distributions (feature size $l \in \{2, 4, 6, 8\}$) (a) $N = 50$ (b) $N = 100$ (c) $N = 150$ (d) $N = 200$. The sharp turns indicate the clustering structure

Method II (cont')

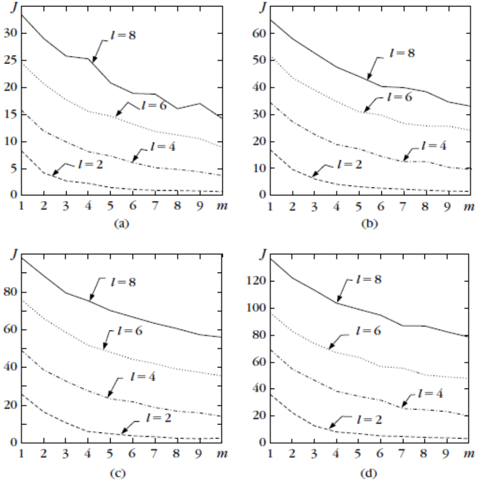

Figure: data set generated from 4 poorly-separated uniformed distributions (feature size $l \in \{2, 4, 6, 8\}$) (a) $N = 50$ (b) $N = 100$ (c) $N = 150$ (d) $N = 200$. No sharp turn exhibited

Hard Clustering Indices

- The modified Hubert Γ statistic: correlation between proximity matrix P and cluster distance matrix Q
  - $P(i,j) = d(x_i, x_j)$, $Q(i,j) = d(c_{x_i}, c_{x_j})$

$$\Gamma = (1/M) \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} X(i,j) Y(i,j) \tag{1}$$

- The Dunn and Dunn-like indices
  - dissimilarity function between two clusters: $d(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x,y)$
  - diameter of a cluster C: $diam(C) = \max_{x,y \in C} d(x,y)$
  - Dunn index:

$$D_m = \min_{i=1,\ldots,m} \min_{j=i+1,\ldots,m} \frac{d(C_i, C_j)}{\max_{k=1,\ldots,m} diam(C_k)} \tag{2}$$

Hard Clustering Indices (Cont')

- The Davies-Bouldin(DB) and DB-like indices:
  - s_i is the measure of the spread around its mean vector for cluster C_i
  - dissimilarity function between two clusters: $d(C_i, C_j)$
  - the similarity index R_{ij} between C_i, C_j has the property:
    - if $s_j > s_k$ and $d_{ij} = d_{ik}$ then $R_{ij} > R_{ik}$
    - if $s_j = s_k$ and $d_{ij} < d_{ik}$ then $R_{ij} > R_{ik}$
  - choose $R_{ij} = \frac{s_i + s_j}{d_{ij}}$, $R_i = \max_{j=1,..m, j \neq i} R_{ij}$

$$DB_m = \frac{1}{m} \sum_{i=1}^{m} R_i \qquad (3)$$

- The DB-like indices based on MST
  - $R_{ij} = \frac{s_i^{MST} + s_j^{MST}}{d_{ij}}$
  - $DB_m^{MST} = \frac{1}{m} \sum_{i=1}^{m} R_i^{MST}$

Hard Clustering Indices (Cont')

- The silhouette index
  - a_i is average distance between x_i and the rest elements of the cluster C_i

$$a_i = d_{avg}^{ps}(x_i, C - x_i) \tag{4}$$

  - b_i is average distance between x_i and its closest cluster C_k

$$b_i = \min_{k=1,\dots,m,k \neq C_i} d_{avg}^{ps}(x_i, C_k) \tag{5}$$

  - the silhouette width of x_i

$$s_i = \frac{b_i - a_i}{\max(b_i, a_i)} \tag{6}$$

- $S_j = \frac{1}{n_j} \sum_{i:x_i \in C_j} s_i$, $S_m = \frac{1}{m} \sum_j^m S_j$

Hard Clustering Indices (Cont')

- The Gap indices:
  - sum of distance between all pairs within the same cluster:
$$D_q = \sum_{x_i \in C_q} \sum_{x_j \in C_q} d(x_i, x_j) \qquad (7)$$
  - $W_m = \sum_{q=i}^{m} \frac{1}{2n_q} D_q$
- for each m, n data set X_m^r, $r = 1, ..., n$ are generated, the estimated size of cluster is obtained by maximizing:

$$Gap_n(m) = E_n(log(W_m^r)) - log(W_m) \qquad (8)$$

Hard Clustering Indices (Cont')

- Information theory based criteria:
  - criteria function

$$C(\theta, K) = -2L(\theta) + \phi(K) \tag{9}$$

  - $L(\theta)$ is the log-likelihood function
  - K is the order of the model - dimentionality of θ, ϕ is an increasing function of K
  - K is strictly increasing function of m

$$K(m, l) = (l + \frac{l(l+1)}{2} + 1)m - 1; \tag{10}$$

- the goal is to minimize C with respect to θ and K

References

S. Theodoridis and K. Koutroumbas. (2009). Pattern Recognition (4th edition), Academic Press.

Questions