Riconoscimento e recupero dell’informazione per bioinformatica

Clustering: validazione

Manuele Bicego

Corso di Laurea in Bioinformatica

Dipartimento di Informatica - Università di Verona

Sommario

- ⇨ Definizione di validazione del clustering
- ⇨ Validazione di gerarchie
- ⇨ Validazione di partizioni
- ⇨ Validazione di cluster
- ⇨ “Clustering tendency”

Definizione

- Validazione del clustering: insieme di procedure che valutano il risultato di un'analisi di clustering in modo quantitativo e oggettivo
  - Differente dalla validazione "soggettiva": data dal particolare contesto applicativo, con l'utilizzo della conoscenza a priori sul problema (intesa anche come "interpretazione dei risultati")
  - In questa parte: validazione "oggettiva": misura quantitativa della capacità della struttura trovata di spiegare i dati (indipendentemente dal contesto)

Indici di validità

Gli indici possono essere diversi a seconda della struttura analizzata (del tipo di clustering)

- Gerarchie: risultato degli algoritmi gerarchici
  - Possiamo anche voler valutare una gerarchia esistente, ad esempio un modello teorico
- Partizioni: risultato degli algoritmi partizionali
  - Si può valutare una partizione esistente derivante da informazioni di categoria
- Clusters: sottoinsiemi di patterns
  - Derivanti da cluster analysis, informazione di categorie, …

Indici di validità

Tipi di indici:

- ⇨ Criteri esterni:
  - ⇨ misurano le performance di un clustering andando a confrontare informazioni a priori
  - ⇨ Esempio: etichette già note a priori
- ⇨ Criteri interni:
  - ⇨ Misurano le performance di un clustering utilizzando solo i dati (completamente non supervisionato)
- ⇨ Criteri relativi:
  - ⇨ Confronta due risultati di clustering

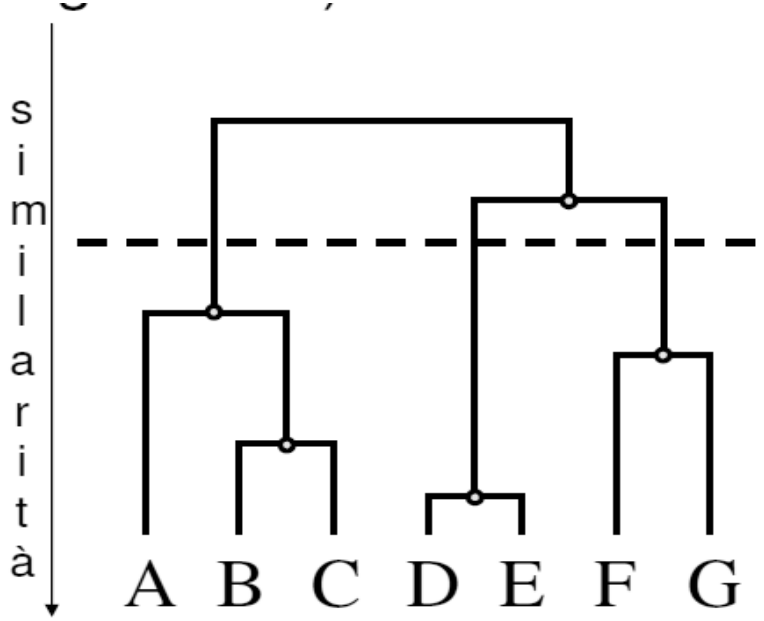
Indici di validità per gerarchie

- Criteri esterni: verificare se una gerarchia (dendrogramma) calcolata per un dato insieme di dati corrisponde alla gerarchia attesa
- Approccio tipico (Hubert's Γ statistics)

- Nota: questo problema di validazione non ha ricevuto un grande interesse, in quanto è piuttosto difficile avere una gerarchia "vera" con cui confrontare il clustering

Criteri interni per gerarchie

- Rispondono alle seguenti domande:
  - Una gerarchia fitta bene i dati su cui è stata calcolata?
  - Ci si può fidare di un determinato risultato di clustering gerarchico?
- Un esempio: CPCC (Cophenetic correlation coefficient)
  - cophenetic distance: il livello di un dendrogramma dove due oggetti sono stati messi nello stesso cluster per la prima volta



d(D,A) = 6

d(D,E) = 1

Criteri interni per gerarchie

- la cophenetic distance misura quando sono simili due oggetti “dato l’albero” (cioè la misura di distanza espressa dall’albero)
- CPCC: coefficiente di correlazione normalizzato

$$CPCC = \frac{\frac{1}{M}\sum_{i,j} d(i,j)d_C(i,j) - m_D m_C}{\left[\frac{1}{M}\sum_{i,j} d^2(i,j) - m_D\right]^{1/2}\left[\frac{1}{M}\sum_{i,j} d_C^2(i,j) - m_C\right]^{1/2}} \qquad (i,j) : 1 \le i < j \le n$$

$$m_D = \frac{1}{M}\sum_{i,j} d(i,j)$$

$$m_C = \frac{1}{M}\sum_{i,j} d_C(i,j)$$

- misura la correlazione tra la distanza derivante dai dati e la distanza derivante dal dendrogramma che spiega i dati
- CPCC varia tra -1 e 1: più è vicino a 1 migliore è il clustering

Indici di validità per partizioni

- Rispondono alle seguenti domande:
  - La partizione ha un buon match con le categorie?
  - Quanti cluster ci sono nel dataset?
  - Dove deve essere tagliato il dendrogramma?
  - Quale tra due partizioni date fitta meglio il dataset?

Indici di validità per partizioni

Criteri esterni:

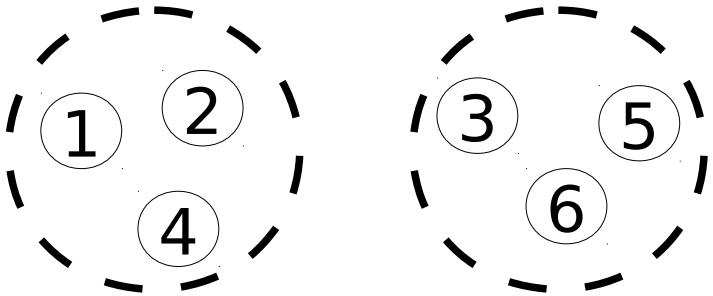
- Tipicamente si va a confrontare due partizioni:
  - Una deriva dal clustering
  - Una deriva dall'informazione a priori (etichette)

- Diversi indici Rand, Jaccard, Fowlkes and Mallows, Γ statistic

Indici di validità per partizioni

- Punto di partenza: una funzione indicatrice $I_U(i,j)$
  - $I_U(i,j)$ vale 1 se gli oggetti i e j sono nello stesso cluster secondo il clustering U

Partizione U



Funzione Indicatrice

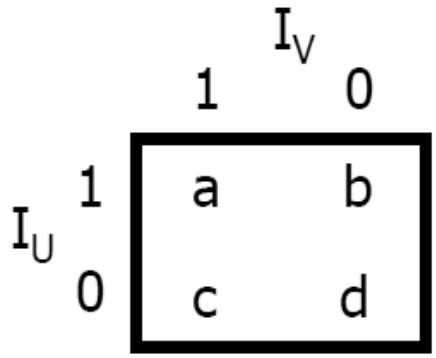
$\bullet I_U$	1	2	3	4	5	6
1	1	1	0	1	0	0
2	1	1	0	1	0	0
3	0	0	1	0	1	1
4	1	1	0	1	0	0
5	0	0	1	0	1	1
6	0	0	1	0	1	1

Indici di validità per partizioni

Tipicamente si hanno due partizioni U e V

- ⇨ U: risultato del clustering
- ⇨ V: clustering “vero” (deriva dalle etichette note a priori)

Posso calcolare la matrice di contingenza



a = numero di coppie di oggetti che sono messi nello stesso cluster in tutte e due le partizioni

b = numero di coppie di oggetti che sono messi nello stesso cluster da U ma non da V

c = numero di coppie di oggetti che sono messi nello stesso cluster da V ma non da U

d = numero di coppie di oggetti messi in cluster diversi sia da U che da V

Indici di validità per partizioni

Matematicamente

$$a = \sum_{i,\,j} I_U(i,\,j)\, I_V(i,\,j)$$

È uguale a 1 se sia U che V sono 1, cioè se sia U che V mettono gli oggetti xi e xj nello stesso cluster

$$b = \sum_{i,\,j} I_U(i,\,j)\,(1 - I_V(i,\,j))$$

È uguale a 1 se U è 1 e V è 0, quindi se U mette xi e xj nello stesso cluster ma V no

Indici di validità per partizioni

$$c = \sum_{i,j} \left(1 - I_U(i,j)\right) I_V(i,j)$$

$$d = \sum_{i,j} \left(1 - I_U(i,j)\right)\left(1 - I_V(i,j)\right)$$

Si possono anche calcolare le seguenti quantità

- ⇨ m_1 = numero di coppie nello stesso gruppo in U
  - ⇨ $m_1 = a+b$
- ⇨ m_2 = numero di coppie nello stesso gruppo in V
  - ⇨ $m_2 = a+c$
- ⇨ M = numero totale di coppie
  - ⇨ $M = a+b+c+d$

Indici di validità per partizioni

I diversi indici sono definiti a partire da queste quantità:

l'idea generale è quella di misurare quanto vanno d'accordo le due partizioni

$$\frac{a+d}{\binom{n}{2}}$$

Indice RAND

$$\frac{a}{(a+b+c)}$$

Indice Jaccard

$$\frac{Ma-m_1 m_2}{\left(m_1 m_2 (M-m_1)(M-m_2)\right)^{1/2}}$$

Γ statistic

$$\frac{a}{(m_1 m_2)^{1/2}}$$

Fowlkes & Mallows

Indici di validità per partizioni

Criteri interni:

- Difficili da stimare: devono misurare il fitting tra una partizione data e il dataset
- Problema fondamentale: stimare il numero di clusters
- Molti metodi (esempio metodi di model selection per modelli probabilistici)
- Ma molte difficoltà:
  - Stima della baseline (campionamento di molti dataset + stima di un indice interno --- ma quale modello per campionare i dati?)
  - Gli indici interni dipendono strettamente dai parametri del problema:
    - Numero di features, numero di patterns, numero di clusters ...

Un particolare indice

L’indice di Davies-Bouldin (1979)

- ⇨ Inizialmente utilizzato per decidere quando fermare un clustering sequenziale
- ⇨ L’indice viene calcolato al variare del numero di clusters
- ⇨ Il miglior clustering corrisponde al valore minimo

L'indice di Davies Bouldin

DEFINIZIONI

- $\{\mathbf{x}_1, \ldots \mathbf{x}_N\}$ punti da clusterizzare
- $C_1..C_K$: partizione da valutare (insieme dei K clusters, ognuno di cardinalità n_j)

Si possono calcolare il centroide, la variazione intracluster e la variazione tra cluster

$$m_j = \frac{1}{n_j} \sum_{x_i \in C_j} x_i$$ centroide

$$e_j^2 = \frac{1}{n_j} \sum_{x_i \in C_j} (x_i - m_j)^T (x_i - m_j)$$ within cluster variation

$$dm(j,h) = d(m_j, m_h)$$ between cluster variation

L'indice di Davies Bouldin

Passi per calcolare l'indice

- Per ogni coppia di cluster (j,h) si calcola

$$R_{jh} = \frac{e_j + e_h}{dm(j,h)}$$

- Per ogni cluster si calcola

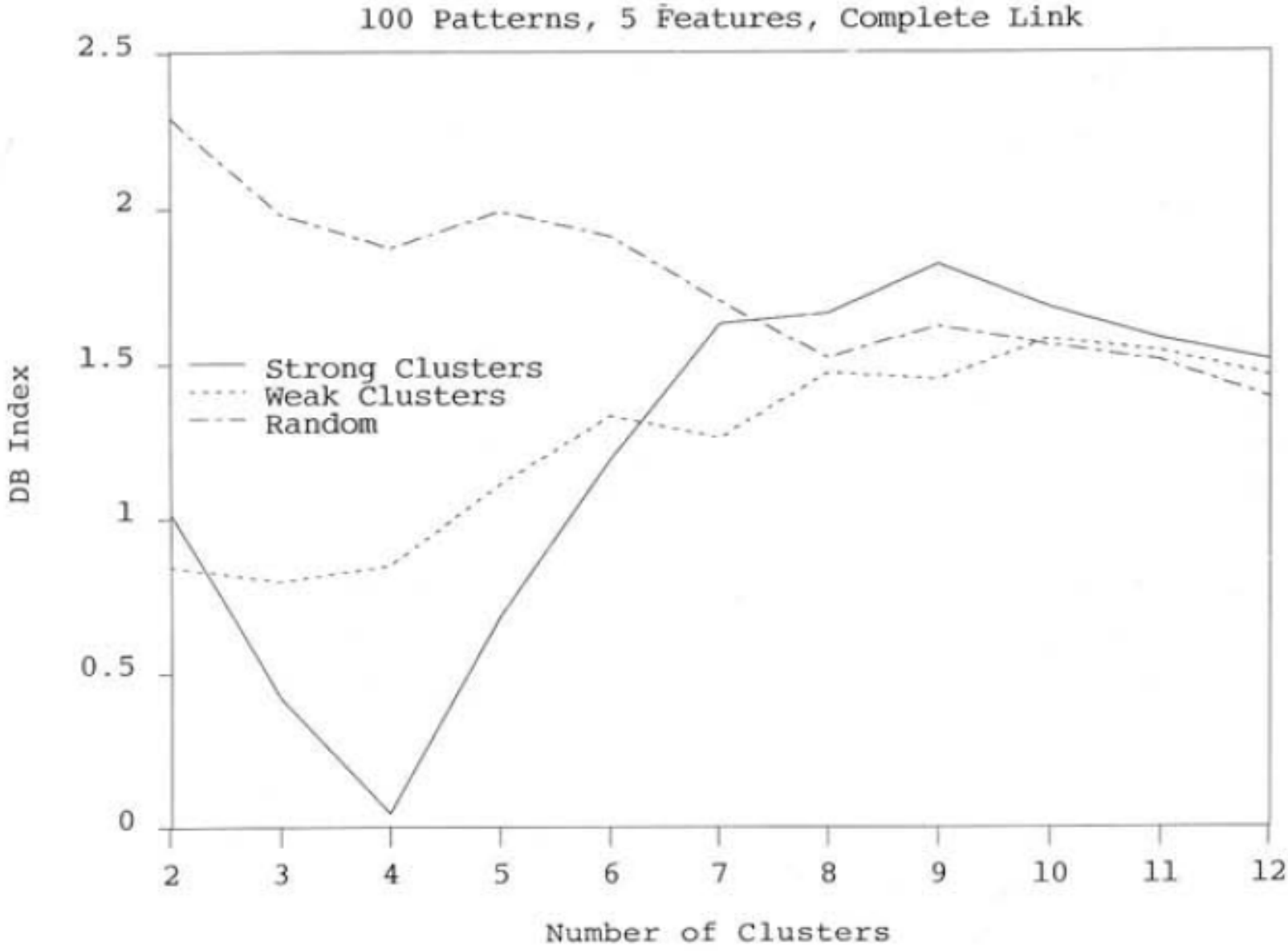
$$R_j = \max_{j \neq h} R_{jh}$$

- L'indice di Davies Bouldin viene determinato come

$$DB(\{C_{1,\ldots}, C_K\}) = \frac{1}{K} \sum_{j=1}^{K} R_j$$

Più piccolo è il valore dell'indice migliore è il clustering!

Può anche essere utilizzato per determinare la presenza di una struttura di clustering



Validità di singoli cluster

- Criteri basati su due proprietà:
  - Compattezza
  - Isolamento
- Compattezza: misura la coesione interna tra gli oggetti del cluster (quanto sono vicini tra di loro)
- Isolamento: misura la separazione tra un cluster e tutti gli altri pattern.

- Cluster valido: compatto e isolato.

Come misurare compattezza e isolamento?

Diversi indici complessi (vedi Cap 4.5 Libro Jain Dubes)

Clustering tendency

- Problema: gli algoritmi di clustering producono sempre un output, indipendentemente dal dataset
- Definizione di cluster tendency: identificare, senza effettuare il clustering, se i dati hanno una predisposizione ad aggregarsi in gruppi naturali

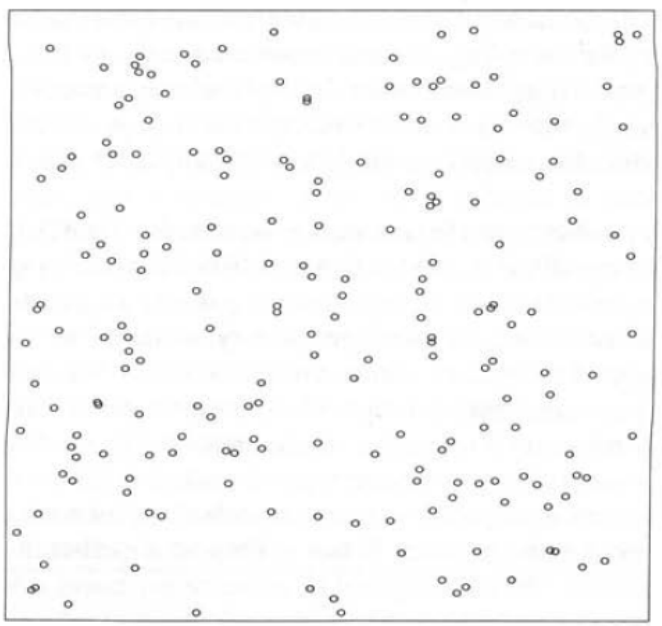
- Operazione preliminare cruciale:
  - Previene dall'applicare elaborate metodologie di clustering e di validazione a dati in cui i cluster sono sicuramente degli artefatti degli algoritmi di clustering

Clustering tendency

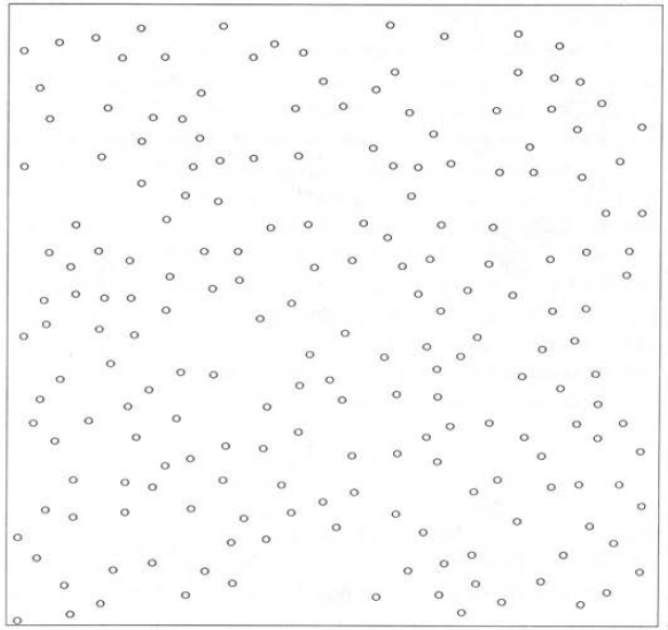
- IDEA: studio dello spazio delle features in modo da identificare tre possibili situazioni:
  1. I pattern sono sistemati in modo casuale (spatial randomness)
  2. I pattern sono aggregati, cioè esibiscono una mutua attrazione
  3. I pattern sono spaziati regolarmente, cioè esibiscono una mutua repulsione

- Nei casi 1 e 3 non ha senso effettuare il clustering

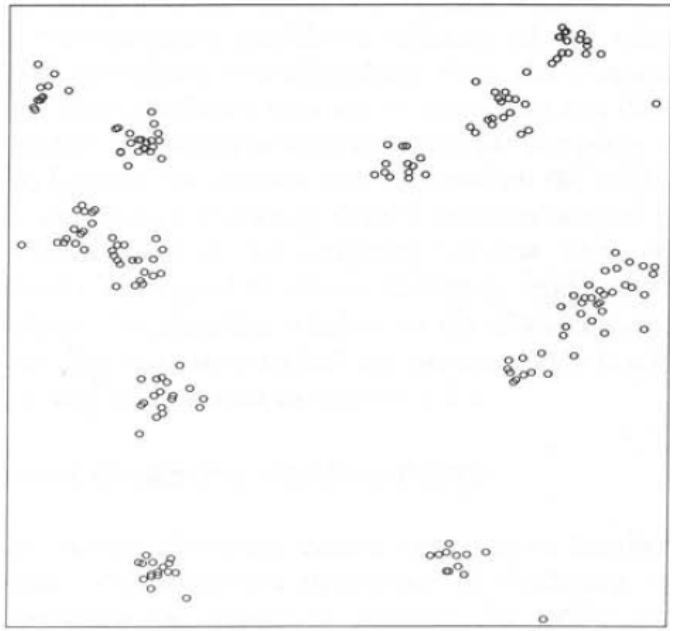
Cluster tendency



random



regular



cluster

Cluster tendency

IDEA: effettuare alcuni test in modo da determinare se esiste o meno una struttura (e.g. test per una distribuzione uniforme in una finestra detta sampling window)

ESEMPI:

- Scan tests:
  - Contare il numero di pattern presenti nella sottoregione più popolosa
  - Se il numero è inusualmente grande allora esiste un clustering
  - PROBLEMI: come definire le sottoregioni, cosa vuol dire “inusualmente grande”