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)

3

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, …

4

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

5

Indici di validità

NOTA: differenza tra criterio e indice

- ⇨ Indice: misura statistica che viene utilizzata per testare la validità
- ⇨ Criterio: strategia con cui un clustering viene validato

TIPICAMENTE:

- ⇨ Nel caso di criteri interni o esterni, si va a vedere se il valore di un indice è particolarmente "grande" o particolarmente "piccolo"

6

Questioni principali

- ⇨ Definizione di un indice:
  - ⇨ Deve avere senso anche da un punto di vista intuitivo
  - ⇨ Deve essere basato su una solida teoria
  - ⇨ Deve essere facilmente calcolabile
- ⇨ Definizione di una baseline
  - ⇨ Distribuzione di base derivante da una popolazione senza struttura
- ⇨ Test per struttura non random
  - ⇨ Come confrontare l'indice con la baseline

7

Indici di validità per gerarchie

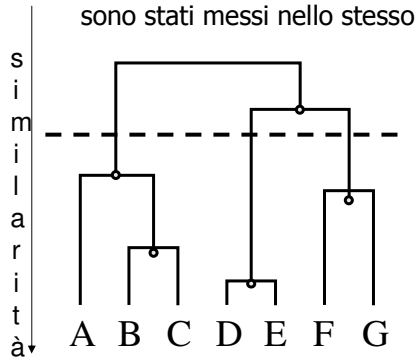
- ⇨ Criteri esterni: verificare se una gerarchia (dendrogramma) calcolata per un dato insieme di dati corrisponde alla gerarchia attesa
- ⇨ Approccio tipico (Hubert's Γ statistics)

- ⇨ 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

8

Criteri interni

- 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

- 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?

11

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

12

Indici di validità per partizioni

- Idea: ho due partizioni U e V da confrontare
  - U: risultato del clustering
  - V: clustering "vero" (deriva dalle etichette note a priori)
- Definizione di due funzioni Indicatrici
  - $I_U(i,j)$ vale 1 se gli oggetti i e j sono nello stesso cluster secondo il clustering U
  - $I_V(i,j)$ vale 1 se gli oggetti i e j sono nello stesso cluster secondo il clustering V
- Definizione della "tabella di contingenza"

I_U \ I_V	1	0
1	a	b
0	c	d

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

13

Indici di validità per partizioni

Altre definizioni

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

14

Indici di validità per partizioni

I diversi indici sono definiti a partire da queste quantità

⇨ Rand $\frac{a+d}{\binom{n}{2}}$

⇨ Jaccard $\frac{a}{a+b+c}$

⇨ Fowlkes & Mallows $\frac{a}{(m_1 m_2)^{1/2}}$

⇨ Γ statistic $\frac{Ma - m_1 m_2}{(m_1 m_2 (M - m1)(M - m2))^{1/2}}$

15

Indici di validità per partizioni

Criteri interni:

⇨ Difficili da stimare: devono misurare il fitting tra una partizione data e il dataset

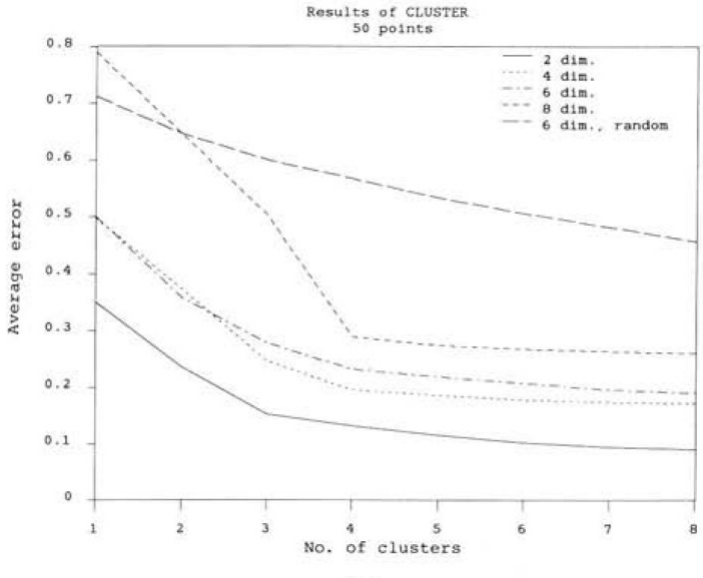
⇨ Problema fondamentale: stimare il numero di clusters

⇨ 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 …

16



⇨ Misura dell'errore medio del k-means

17

Un particolare indice

L'indice di Davies-Bouldin (1979)

⇨ Inizialmente utilizzato per decidere quando fermare un clustering sequenziale

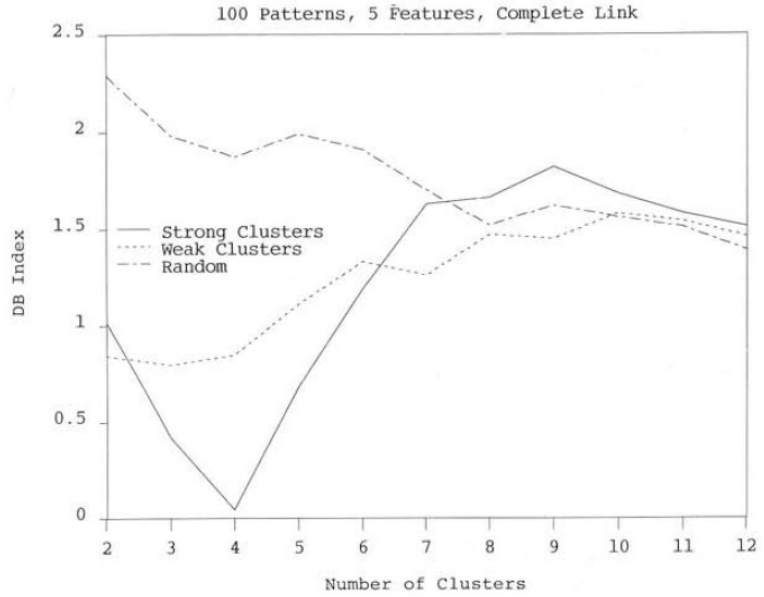
⇨ L'indice viene calcolato al variare del numero di clusters

⇨ Il miglior clustering corrisponde al valore minimo

⇨ (definizione alla lavagna)

18

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)

20

Clustering tendency

- ⇨ Problema: gli algoritmi di clustering producono sempre un output, indipendentemente dal dataset
- ⇨ Definizione: 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

21

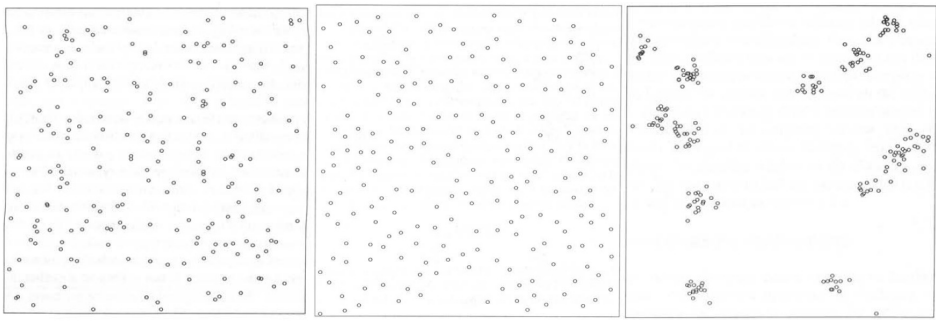
Clustering tendency

- ⇨ 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

22

Cluster tendency



random regular cluster

23

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"

24

Cluster tendency

ESEMPI

⇨ Quadrat analysis:

- ⇨ Partizionare la sampling window in rettangoli di dimensione uguale
- ⇨ Contare il numero di punti in ogni rettangolo
- ⇨ In caso di distribuzione randomica, l'insieme di conteggi segue una distribuzione nota
- ⇨ La randomicità viene attestata con un test di similarità tra distribuzioni (e.g. il test del chi quadro)
- ⇨ PROBLEMI:
  - ⇨ dimensione dei sottorettangoli (si può anche fare multiscala)
  - ⇨ Funziona solo per un numero limitato di features (altrimenti la maggior parte dei rettangoli sono vuoti)

25