

Riconoscimento e recupero dell'informazione per bioinformatica

Clustering: introduzione

Manuele Bicego

Corso di Laurea in Bioinformatica
Dipartimento di Informatica - Università di Verona

Una definizione possibile

[Jain *et al.*, ACM Computing Surveys, 1999]

- ⇒ Il clustering rappresenta l'organizzazione di un insieme di patterns (entità) in gruppi (clusters) sulla base della similarità
 - ⇒ Pattern: entità di interesse, come sequenze di geni, spettri di risonanza, ...
 - ⇒ i pattern in un gruppo sono tutti simili tra loro, i pattern di gruppi diversi sono invece differenti tra di loro
- ⇒ I cluster sono insiemi di pattern simili
- ⇒ Il processo è completamente "non supervisionato"
 - ⇒ Non è data nessuna informazione a priori sui gruppi

Nota

- ⇒ Il termine "data clustering" rappresenta un concetto utilizzato in molte comunità:
 - ⇒ Pattern Recognition, Statistical Data Analysis, Machine Learning, Knowledge and Data Engineering, Psychology, Geology
 - ⇒ In ogni contesto ci sono diverse terminologie, assunzioni, ipotesi
- ⇒ In generale, il significato comune è quello di "metodi per raggruppare dati *non etichettati* (dati di cui non si conosce la categoria/la classe)
- ⇒ In questo corso: il punto di vista della Pattern recognition
 - ⇒ Il più vicino alla bioinformatica / il più utilizzato in questo contesto₃

Intrinsecamente un problema mal posto

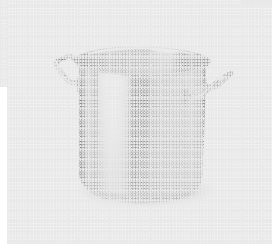
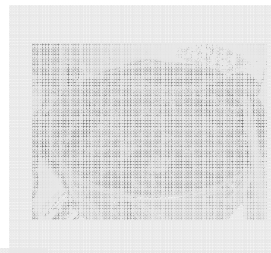
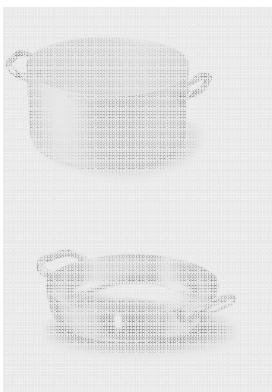
- ⇒ "Il clustering rappresenta l'organizzazione di un insieme di patterns (entità) in gruppi (clusters) sulla base della similarità"
- ⇒ Qual'è la similarità più appropriata?
 - ⇒ Cambiare la similarità cambia il risultato
- ⇒ Cosa deve rappresentare un "buon gruppo"?
 - ⇒ Il concetto di gruppo è definito in modo vago e assolutamente soggettivo
 - ⇒ Il processo è non supervisionato: non sappiamo se facciamo giusto! (differentemente dalla classificazione)

Esempio: Oggetti da clusterizzare



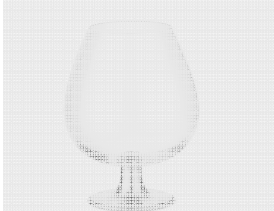
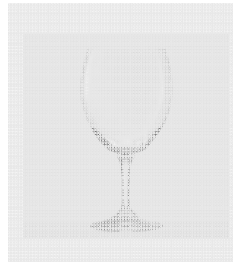
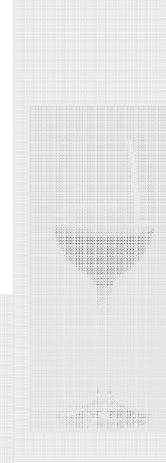
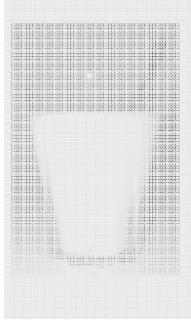
5

Primo gruppo: bicchieri



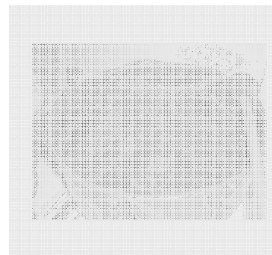
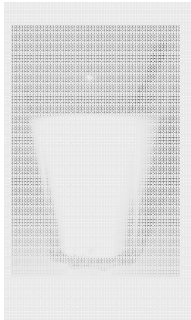
6

Secondo gruppo: padelle



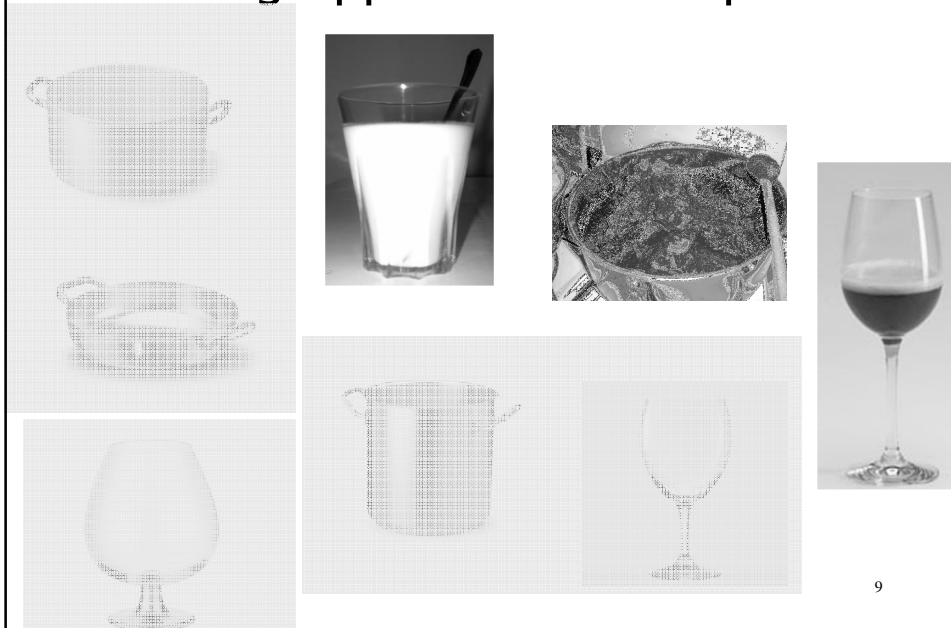
7

Primo gruppo: contenitori vuoti



8

Secondo gruppo: contenitori pieni



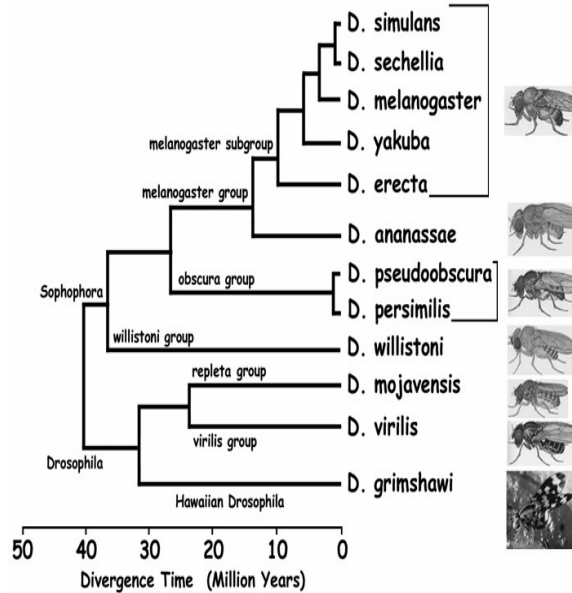
Quindi

- ⇒ Il concetto di cluster è vago:
- ⇒ Dipendentemente dalle misure di similarità utilizzate cambia il risultato
- ⇒ La scelta della misura di similarità è cruciale.
 - ⇒ Dovrebbe essere fatta in modo da inglobare la maggior quantità possibile di informazione a priori.
- ⇒ Il risultato può cambiare anche a seconda della metodologia utilizzata per fare clustering (il concetto sarà più chiaro in seguito)

Esempio guida: la filogenesi

⇒ Filogenesi: inferire le relazioni genealogiche tra gli organismi

⇒ clustering di sequenze geniche o proteiche



Rappresentazione dei Pattern

⇒ Pattern: misure dell'oggetto

⇒ spesso è un vettore d-dimensionale $x = [x_1, x_2, \dots, x_d]$ (dove ogni x_i rappresenta una feature)

⇒ ogni pattern è un punto in uno spazio d-dimensionale

⇒ Esempi: misure fisiche dell'oggetto, immagini, spettri

⇒ Spesso definiti come "dati grezzi"

⇒ Altre possibili rappresentazioni:

⇒ stringhe (exe DNA)

⇒ dati tridimensionali (exe: proteine foldate (ripiegate))

Rappresentazione dei Pattern

- ⇒ Pre-processing dei dati: "migliorare" la rappresentazione:
 - ⇒ ridurre il numero dello spazio delle features (per visualizzare, per ridurre il carico computazionale, ...)
 - ⇒ mettere in evidenza particolari strutture o migliorare le capacità discriminative dello spazio
- ⇒ Estrazione di feature: trasformazione dello spazio originale (PCA, ICA, LDA, MDS,...)
- ⇒ Selezione di features: selezionare le feature migliori (SFFS, SBFS, Matching Pursuit, ...)

15

Esempio

- ⇒ Insetti da clusterizzare



- ⇒ Dati grezzi: le sequenze di DNA relative ad un determinato gene

```
CAGATCTTGACGATCCCAAGTGGTTCATTGGCTTTAGATGAAG
TACCGATCTTGACGATCCCAAGTTCATTGGCTTTAGATGAAG
CAGATCTTGACGATCCCAAGTGGTTCATTGGCTTTAGATGAAG
CAGATCTTCACGATCCCAAGTGGTTCATTGGCTTTAGAT
```

- ⇒ Pre-processing: allineamento delle sequenze

```
CA--GATCTTGACGATCCCAAGTGGTTCATTGGCTTTAGATGAAG
TACCGATCTTGACGATCCCAAG----TCATTGGCTTTAGATGAAG
CA--GATCTTGACGATCCCAAGTGGTTCATTGGCTTTAGATGAAG
CA--GATCTTCACGATCCCAAGTGGTTCATTGGCTTTAGAT----
```

16

Similarità

- ⇒ Il concetto di similarità è strettamente incapsulato nella definizione di cluster
 - ⇒ la maggior parte degli algoritmi di clustering dipendono strettamente dalla definizione di questa misura
- ⇒ Esistono molte definizioni diverse
 - ⇒ dipendentemente dal dominio
 - ⇒ dipendentemente dal tipo di feature
 - ⇒ dipendentemente dalla conoscenza a priori
- ⇒ similarità / distanza

17

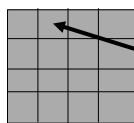
Esempio

- ⇒ Misura di similarità:
 - ⇒ la distanza tra due sequenze è rappresentata dal numero di "sostituzioni" che ci sono, pesate in un certo modo
- ESEMPIO:
 - ⇒ misura di Jukes-Cantor (p = proporzione di nucleotidi dove le due sequenze differiscono)

$$d(S_1, S_2) = -\log\left(1 - \frac{4}{3}p\right)$$

- ⇒ Si calcola la distanza tra tutte le coppie di sequenze

Matrice di
distanza
4x4



Distanza tra la prima e la
seconda sequenza

18

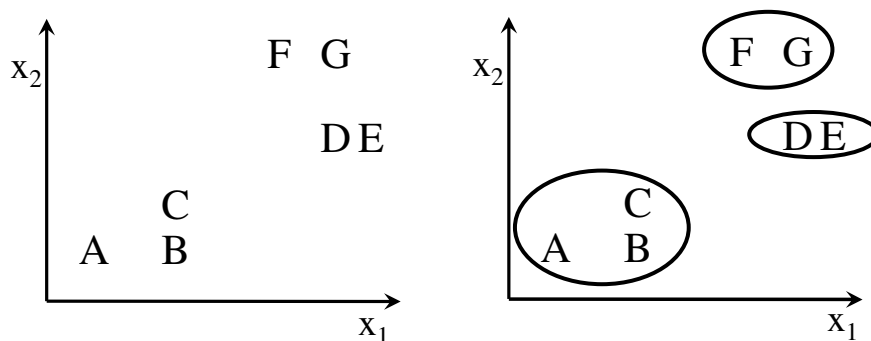
Metodologie di Clustering

- ⇒ Obiettivo: trovare i gruppi data la definizione di similarità
- ⇒ Non esiste un'unica metodologia appropriata per tutti i problemi
 - ⇒ la scelta di un algoritmo appropriato dipende dal dominio, dal processo di acquisizione, dalla conoscenza a priori, dalla quantità di dati a disposizione
- ⇒ Ci sono molti metodi in letteratura
 - ⇒ Diversi criteri di ottimizzazione, assunzioni, modelli, requisiti computazionali
- ⇒ Principale suddivisione: metodi partizionali o gerarchici

19

Metodi partizionali

il risultato è una singola partizione del dataset (il numero di cluster è dato a priori)

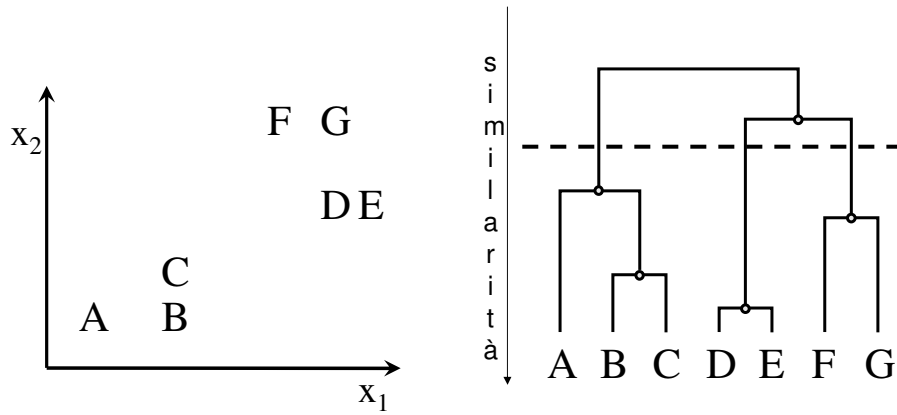


Esempi: K-means (e le sue varianti), ISODATA, PAM, ...

20

Metodi gerarchici

il risultato è una serie di partizioni innestate (un albero binario detto "dendrogramma")



Esempi: Complete Link, Single Link, Ward, ...

21

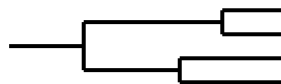
Esempio

⇒ Clustering

4x4  Similarity matrix



Clustering gerarchico



Albero filogenetico

22

Validazione del clustering

- ⇒ Ogni algoritmo di clustering genera SEMPRE un risultato
- ⇒ Approcci differenti tipicamente portano a differenti clusters
- ⇒ Non c'è il "ground truth", il processo è non supervisionato



La validazione dei cluster è fondamentale

- ⇒ Domande:
 - ⇒ I dati sono casuali o esiste qualche giustificazione per il clustering?
 - ⇒ I cluster che determino sono ottimali? E in che senso?

23

Esempio

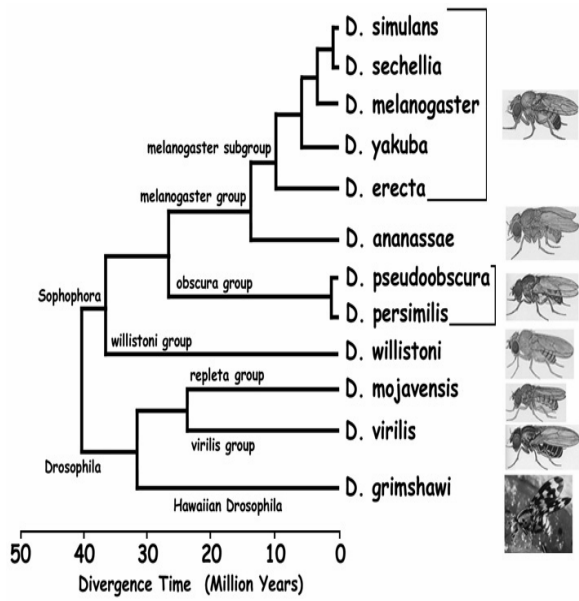
- ⇒ Validazione del clustering

Analisi della robustezza del clustering: BOOTSTRAP

- ⇒ Vengono creati N nuovi data set (per esempio 1000) campionando casualmente N colonne (con rimpiazzo)
 - ⇒ in questo modo in ogni dataset generato contiene lo stesso insieme di specie, con alcuni dei nucleotidi duplicati e con altri rimossi
- ⇒ Per ogni data set viene costruito l'albero (clustering)
- ⇒ Viene calcolata la frequenza con cui ogni sottogruppo dell'albero viene ripetuta
- ⇒ Questa indica la robustezza di un raggruppamento

24

Esempio



⇒ D. simulans and D. sechellia sono più simili tra di loro che rispetto agli altri

⇒ divergenza evolutiva più recente