Riconoscimento e recupero dell'informazione per bioinformatica

Clustering: similarità

Manuele Bicego

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

Sommario

- Definizioni preliminari
- Similarità tra punti
- Similarità tra sequenze
  - dynamic time warping
- Similarità tra insiemi
- Un esempio biologico: il BLAST

Definizioni

⇨ Coefficiente di similarità:
- ⇨ indica la "forza" della relazione tra due oggetti
- ⇨ maggiore è la somiglianza tra questi oggetti, più alto è il coefficiente di similarità

⇨ Dissimilarità (distanza):
- ⇨ concetto simile ma che misura le differenze tra due oggetti

⇨ In generale si può parlare di "misure di prossimità"

$$\mathcal{X} \text{ Dominio del problema }, \mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}$$
$$proximity(\mathbf{x}_i, \mathbf{x}_j) = f : \mathcal{X}^2 \rightarrow \Re$$

3

Definizioni

⇨ Concetto di "metrica": misura di prossimità con particolari caratteristiche

⇨ Definizione: (dissimilarità/distanza metrica):
- ⇨ misura di dissimilarità che soddisfa alle seguenti proprietà:

Non negativitá	$f(\mathbf{x}_i, \mathbf{x}_j) \geq 0$
Riflessivitá	$f(\mathbf{x}_i, \mathbf{x}_j) = 0 \iff \mathbf{x}_i = \mathbf{x}_j$
Commutativitá	$f(\mathbf{x}_i, \mathbf{x}_j) = f(\mathbf{x}_j, \mathbf{x}_i)$
Disuguaglianza triangolare	$f(\mathbf{x}_i, \mathbf{x}_j) \leq f(\mathbf{x}_i, \mathbf{x}_k) + f(\mathbf{x}_j, \mathbf{x}_k)$

4

Altra rappresentazione

⇨ Matrice di prossimità:

⇨ matrice che descrive i valori della funzione per tutte le possibili coppie

$$\begin{pmatrix} 0 & d_{12} & \cdots & d_{1n} \\ d_{21} & 0 & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & 0 \end{pmatrix}$$

⇨ simmetrica / asimmetrica, dissimilarità / similarità, ...

5

Commenti

⇨ La scelta della misura è cruciale e influenza enormemente l'uscita del clustering

⇨ Informazione a priori:

⇨ contesto applicativo

⇨ tipo di pattern (vettore, sequenza, dati mancanti)

⇨ dimensionalità del pattern

⇨ scala

⇨ cardinalità dell'insieme

⇨ requisiti (velocità vs precisione): e.g. retrieval by content

⇨ (esperienza del ricercatore)

6

Trasformazione

⇨ Similarità e dissimilarità misurano la stessa quantità da due punti di vista differenti

⇨ Trasformazione: $d(\mathbf{x}_i, \mathbf{x}_j)$ una distanza

$$s(\mathbf{x}_i, \mathbf{x}_j) = \frac{a}{d(\mathbf{x}_i, \mathbf{x}_j)} \qquad \text{con } a > 0$$

$$s(\mathbf{x}_i, \mathbf{x}_j) = d_{max} - d(\mathbf{x}_i, \mathbf{x}_j) \qquad \text{con } d_{max} = \max_{i,j} = d(\mathbf{x}_i, \mathbf{x}_j)$$

$$s(\mathbf{x}_i, \mathbf{x}_j) = \ln(d_{max} + k - d) \qquad \text{con } k > 0$$

$$s(\mathbf{x}_i, \mathbf{x}_j) = \frac{kd}{1+d} \qquad \text{con } k > 0$$

Misure per pattern vettoriali

⇨ Campo molto investigato: esistono molte misure diverse!
- ⇨ vettori numerici
- ⇨ vettori categorici
- ⇨ vettori binari

⇨ Distanze tra vettori numerici
- ⇨ distanza euclidea
- ⇨ distanza di Manhattan
- ⇨ distanza Maximum
- ⇨ distanza di Mahalanobis
- ⇨ distanza di Minkowski
- ⇨ misura coseno (similarità)

⇨ concetto di distanza "geodesica"

8

Vettori numerici

Nozioni preliminari: vettori **x** = [x_1…x_d], **y** = [y_1…y_d]

⇨ distanza euclidea

$$d(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{j=1}^{d}(x_j - y_j)^2} = [(\mathbf{x}-\mathbf{y})(\mathbf{x}-\mathbf{y})^T]^{1/2}$$

⇨ Molto utilizzata

⇨ distanza di Manhattan (city block distance)

$$d(\mathbf{x},\mathbf{y}) = \sum_{j=1}^{d}|x_j - y_j|$$

⇨ Tutti i percorsi più brevi hanno la stessa lunghezza

⇨ Utilizzata nei circuiti dove i fili possono andare solo orizzontalmente o verticalmente

9

Vettori numerici

⇨ Minimum distance (distanza "sup")

$$d(\mathbf{x},\mathbf{y}) = \max_{1\le j\le d}|x_j - y_j|$$

⇨ Distanza di Mahalanobis

$$d(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})\Sigma(\mathbf{x}-\mathbf{y})^T}$$

⇨ Scalamento degli assi

⇨ Pro: invariante alle rotazioni/traslazioni/trasformazioni affini

⇨ Contro: calcolo della matrice di covarianza

10

Vettori numerici

⇨ Distanza di Minkowsky

$$d(\mathbf{x},\mathbf{y}) = \left(\sum_{j=1}^{d}(x_j - y_j)^p\right)^{1/p}$$

⇨ Generalizzazione della distanza euclidea (p=2) e di quella di manhattan (p=1)

⇨ Similarità coseno

$$d(\mathbf{x},\mathbf{y}) = \frac{<\mathbf{x},\mathbf{y}>}{||\mathbf{x}||||\mathbf{y}||}$$

⇨ Similarità (non distanza)

⇨ Tiene conto della lunghezza dei vettori

11

Misure per dati categorici

⇨ Dati discreti (exe DNA)

⇨ simple matching dissimilarity measure

$$\delta(x,y) = \begin{cases} 0 & se \quad x = y \\ 1 & se \quad x \neq y \end{cases}$$

$$d(\mathbf{x},\mathbf{y}) = \sum_{j=1}^{d}\delta(x_j, y_j)$$

⇨ Misure per dati binari

⇨ Hamming distance: numero di posizioni dove i due vettori differiscono

$$d(\mathbf{x},\mathbf{y}) = \sum_{j=1}^{d}(x_j - y_j)^2$$

12

La distanza di Hamming tra 1011101 e 1001001 è 2

Misure per dati binari

Dati binari: 0 o 1

⇨ Distanza di Hamming: numero di posizioni dove i due vettori differiscono

$$d(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{d} (x_j - y_j)^2$$

⇨ Esempio: la distanza di Hamming tra 1011101 e 1001001 è 2

⇨ Distanza di Jaccard: misura del grado di overlap fra 2 insiemi A e B

⇨ L'intersezione di A e B divisa per l'unione di A e B

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

13

Misure per dati binari

⇨ Versione per dati binari asimmetrici:

⇨ Date 2 stringhe A e B:

$$J' = \frac{M_{01} + M_{10}}{M_{01} + M_{10} + M_{11}}$$

⇨ M_{11} n. di simboli dove sia A che B valgono 1.

⇨ M_{01} n. di simboli dove A vale 0 e B vale 1.

⇨ M_{10} n. di simboli dove A vale 1 e B vale 0.

⇨ Molto utilizzata in bioinformatica

14

Misure per pattern sequenziali

- Devono gestire:
  - l'ordinalità dei dati
  - il fatto che ci possono essere sequenze di lunghezza diversa
- Classi di misure
  - misure per dati vettoriali (se le sequenze sono della stessa lunghezza)
  - misure basate sul concetto di MSC (Massima sottosequenza comune)
  - misure basate su modelli probabilistici
  - misure basate su il concetto di editing (Edit Distance)

15

Misure per pattern sequenziali

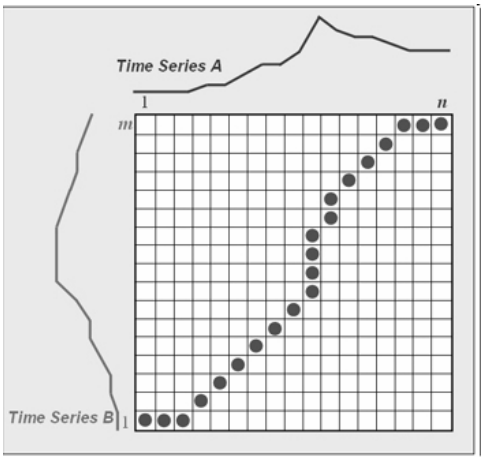
- misure per dati vettoriali (se le sequenze sono della stessa lunghezza)
  - non tengono conto della sequenzialità dei dati
  - non possono gestire sequenze a lunghezza diversa
- Possibile soluzione: dynamic time warping

- Dynamic time warping: metodo per gestire "accelerazioni" e "decelerazioni" della sequenza
  - usato in speech processing
  - GOAL: "allineare" due sequenze "deformando" (warping) l'asse temporale fino a quando non si trova un match ottimale.
- IDEE:
  - estendere le sequenze ripetendo i simboli
  - calcolare la distanza tra le sequenze estese

16

Misure per pattern sequenziali

- ⇨ IDEA: warping path:
  - ⇨ come allineare diversi punti delle due sequenze
  - ⇨ ogni punto rappresenta un matching tra due punti delle due sequenze
  - ⇨ ci possono essere più "match" per lo stesso punto
- ⇨ Dettagli (alla lavagna)



17

Misure per pattern sequenziali

- ⇨ Misure basate sul concetto di MSC (Massima sottosequenza comune)
- ⇨ MSC without scaling:
  - ⇨ IDEA: due sequenze sono simili se hanno una sottosequenza in comune molto grande
  - ⇨ Approcci:
    - ⇨ caso discreto: classico problema della massima sottosequenza comune (vedi corso di algoritmi)
    - ⇨ caso continuo: occorre definire quando due valori "corrispondono" (match)
    - ⇨ ESEMPIO: $x_i \text{ matches } y_j \quad \text{se} \quad |x_i - y_j| < \delta$
  - ⇨ Tipica soluzione con Programmazione Dinamica

18

Misure per pattern sequenziali

⇨ MCS con Local Scaling

⇨ IDEA: due sequenze sono simili se hanno un numero "sufficientemente elevato" di coppie di sottosequenze simili

⇨ due sottosequenze sono simili se la prima può essere scalata e traslata appropriatamente in modo da assomigliare all'altra

⇨ SOLUZIONI: R-tree, percorso più lungo in un grafo aciclico

19

Misure per pattern sequenziali

⇨ misure basate su modelli probabilistici:

⇨ modellare la sequenza con modelli probabilistici

⇨ estrarre la similarità sfruttando i modelli

⇨ ESEMPIO: similarità tra sequenze con Hidden Markov Models

⇨ Hidden Markov Models (HMM)

⇨ Modello statistico per sequenze molto utilizzato in svariati ambiti

20

Misure per pattern sequenziali

⇨ Calcolo della matrice di similarità

⇨ Costruire un modello λ_i per ogni sequenza O_i

⇨ Definire la similarità come:

$S=S(O_i, O_j)$

$S(O_i, O_j)$ = funzione della likelihood

ESEMPIO.

$$S(\mathbf{O}_i, \mathbf{O}_j) = \frac{S_{ij} + S_{ji}}{2}$$

dove

$$S_{ij} = log(P(\mathbf{O}_i|\lambda_j))$$

21

Misure per pattern sequenziali

⇨ Edit distance:

⇨ distanza che tiene in considerazione l'ordine con cui un pattern sequenziale è composto

⇨ ESEMPIO: i simboli sono lettere, i pattern sono parole di un testo scritto.

⇨ Molto utilizzato per automatic editing e text retrieval (trovare il best match tra un pattern e un database di patterns)

⇨ Utilizzabile per trovare la distanza tra due sequenze di geni

⇨ Errori che deve considerare:

⇨ cambiamenti: "befuty" invece di "beauty"

⇨ inserzioni: "beazuty"

⇨ cancellazioni: "beuty"

22

Misure per pattern sequenziali

- Filosofia: "variational similarity"
- La similarità tra due pattern è basata sul "costo" che devo pagare per convertire un pattern nell'altro
- Edit distance

$$D(A,B) = min_j[C(j) + I(j) + R(j)]$$

- j varia tra tutte le possibili variazioni necessarie per ottenere B da A
- Soluzione algoritmica: programmazione dinamica (dettagli nel cap 8.2.2 del Theodoridis)

23

Misure per insiemi

- Obiettivo: trovare una misura di prossimità per insiemi.
  - Dati non ordinati
  - cardinalità diversa
- Altro punto di vista:
  - misure di similarità tra clusters (i clusters sono insiemi)
  - utilizzabile negli algoritmi gerarchici
- ESEMPI
  - nearest neighbor distance: distanza tra i punti più vicini
  - farthest neighbor distance: distanza tra i punti più lontani
  - average neighbor distance: distanza media tra tutti i punti
  - mean based distance: distanza tra i "rappresentanti" dei cluster
    - Media
    - Medoide
    - Vettore più centrale

24

Un esempio biologico: BLAST

- Basic Local Alignment Search Tool
  - algoritmo per confrontare sequenze biologiche (nucleotidiche o aminoacidiche)
  - confronta una sequenza di test con un database di sequenze, ritornando le più simili
  - uno degli algoritmi più famosi di bioinformatica
    - affronta un problema molto importante
    - è computazionalmente efficiente -- la ricerca effettuata con algoritmi di programmazione dinamica è assolutamente inefficiente, vista la mole di dati presente oggigiorno – BLAST è 50 volte più veloce

25

Un esempio biologico: BLAST

- IDEA: cerca di allineare due sequenze, lo score di allineamento rappresenta la misura della bontà del match

- Assunzioni / Idee per velocizzare l'approccio
  - non cercare l'allineamento "ottimale"
  - non effettuare la ricerca in tutto lo spazio delle sequenze
  - utilizzare una serie di euristiche per velocizzare l'approccio

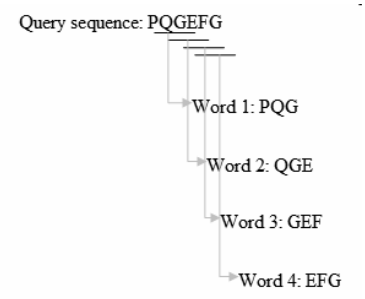
- Input dell'algoritmo:
  - sequenza query (sequenza sconosciuta)
  - sequenza target (o database)

26

Un esempio biologico: BLAST

PASSI dell'ALGORITMO

1. Rimuovere le regioni di bassa complessità della sequenza query
   - ⇨ regioni della sequenza con una ripetizioni di pochi tipi di simbolo
   - ⇨ possono confondere il programma nello trovare regioni significative
2. Creare una lista delle "word" di K lettere della sequenza query



27

Un esempio biologico: BLAST

3. cercare, in tutte le sequenze del database, tutte le word di lunghezza K che hanno un buon match con le word della sequenza query
   - ⇨ buon match = score di allineamento sopra una certa soglia
   - ⇨ utilizzo della "substitution matrix" per calcolare lo score
   - ⇨ lo score considera l'allineamento senza gap
   - ⇨ ogni word trovata si chiama "hit" (o "hotspot")
   - ⇨ allineamento senza gap è molto veloce: possibilità di memorizzare una volta per tutte le posizioni delle word in tutto il database
4. utilizzare ogni "hit" come "seme" per allargare la regione di similarità
   - ⇨ cercare di estendere la coppia di similarità a dx e a sx fino a quando lo score di similarità non diminuisce
   - ⇨ il risultato si chiama HSP (High Scoring segment pair)

28

Un esempio biologico: BLAST

5. visualizzare tutti gli HSP con uno score sufficientemente alto
   - ⇨ vengono listati in ordine di score
6. fornire un'analisi statistica degli score risultanti: l'E-value
   - ⇨ misura il numero di hit che si potrebbero vedere "per caso", in un database di sequenze casuali
   - ⇨ dipende dalla dimensionalità del database e dalla lunghezza della sequenza di query
   - ⇨ la significatività statistica è proporzionale al valore di tale indice (valori attorno allo zero supportano fortemente i risultati)

29

Un esempio biologico: BLAST

Note:

- ⇨ Eventualmente si può gestire anche la presenta di più HSP in una stessa sequenza del database
- ⇨ si può utilizzare on line: http://blast.ncbi.nlm.nih.gov/Blast.cgi
- ⇨ utilizzatissimo per il buon compromesso tra accuratezza e velocità (negli anni sono state presentate molte varianti)
  - ⇨ l'articolo dove viene presentato è il più citato degli anni 90

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). "Basic local alignment search tool". *J Mol Biol* **215** (3): 403–410

30