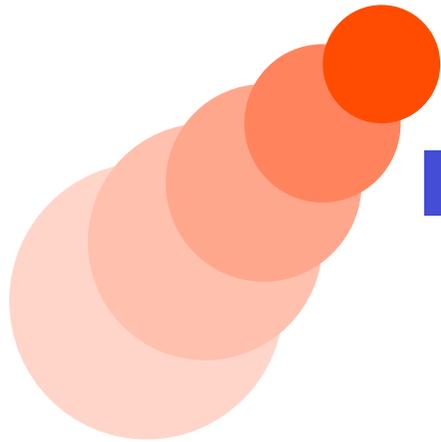


Data Warehousing, Data Mining & Business Intelligence



Clustering

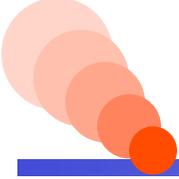
Paolo G. Franciosa

Dipartimento di Statistica, Probabilità e Statistiche Applicate

Università "La Sapienza"

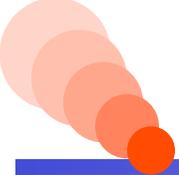
paolo.franciosa@uniroma1.it

Questo materiale deriva dalla traduzione e adattamento delle presentazioni pubblicate dal prof. Jiawei Han



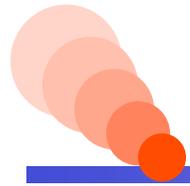
Cluster Analysis

- **Introduzione**
- Tipi di dato nella cluster analysis
- Approcci
- Metodi basati sul partizionamento
- Metodi gerarchici
- Metodi basati sulla densità
- Scoperta di outlier



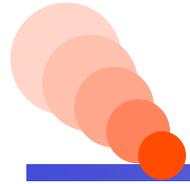
Cos'è la Cluster Analysis?

- Cluster: un insieme di oggetti
 - simili a quelli dello stesso cluster
 - dissimili da quelli in altri cluster
- Cluster analysis
 - raggruppare oggetti in cluster
- Può essere vista come una tecnica di **classificazione non supervisionata** (classi non definite a priori)
- Applicazioni tipiche
 - strumento **stand-alone** per studiare la distribuzione dei dati
 - trattamento **preliminare** per altri algoritmi



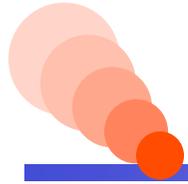
Applicazioni generali

- Pattern Recognition
- Analisi di dati spaziali
 - mappe tematiche in GIS
 - individuare cluster spaziali e la loro spiegazione (mining)
- Image Processing
- Scienze economiche (ricerche di mercato, segmentazione)
- WWW
 - classificazione di documenti
 - analisi di dati di log per individuare gruppi dal comportamento omogeneo



Esempi di applicazione

- Marketing: individuare gruppi di clienti per programmi di marketing mirati
- Territorio: osservazione della terra – aree di utilizzo omogeneo
- Assicurazioni: identificare gruppi di assicurati dal costo elevato
- Pianificazione urbana: identificare gruppi di edifici in base a tipo, valore, posizione
- Studio di terremoti: individuazione di faglie in base alla posizione degli epicentri

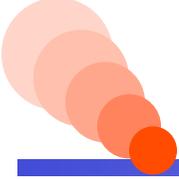


Caratteristiche di un buon clustering

- Un clustering di buona qualità dovrebbe produrre
 - elevata similarità intra-cluster
 - bassa similarità inter-cluster

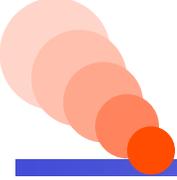
- Punti cruciali
 - misura di similarità
 - implementazione
 - valutazione del clustering raggiunto (per metodi iterativi)

- Capacità di scoprire pattern “nascosti”



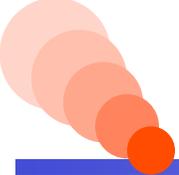
Requisiti di un algoritmo di clustering per Data Mining

- Scalabilità
- Applicabilità a tipi diversi di attributi
- Cluster di forma arbitraria
- Numero di cluster non prefissato
- Parametri di input determinabili senza conoscenza approfondita del dominio applicativo
- Capacità di gestire rumore e outliers
- Insensibilità rispetto all'ordine dei dati di input
- Numero di dimensioni elevato
- Interpretabilità e utilizzabilità dei risultati



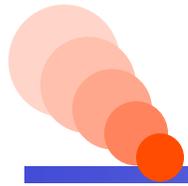
Cluster Analysis

- Introduzione
- **Tipi di dato nella cluster analysis**
- Approcci
- Metodi basati sul partizionamento
- Metodi gerarchici
- Metodi basati sulla densità
- Scoperta di outlier



Misure

- La (dis)similarità è espressa attraverso una funzione “distanza” $d(i, j)$
- La “bontà” di un clustering è espressa attraverso una funzione diversa, spesso soggettiva
- Definizione di distanza:
 - dati numerici, su varie scale
 - boolean/binari
 - categorici
 - ordinali
 - altro (multimediali, ...)
- Necessità di definire criteri di pesatura sui diversi attributi
- Misure necessariamente soggettive



Dati numerici

- Normalizzazione dei dati

- deviazione assoluta dalla media:

$$s_f = \frac{1}{n} (|x_{1f} - m_f| + |x_{2f} - m_f| + \dots + |x_{nf} - m_f|)$$

dove $m_f = \frac{1}{n} (x_{1f} + x_{2f} + \dots + x_{nf})$.

- valore normalizzato (*z-score*)

$$z_{if} = \frac{x_{if} - m_f}{s_f}$$

- La deviazione assoluta è più robusta della deviazione standard

Misure di similarità per dati numerici

- Una classe di distanze: *Minkowski distance*:

$$d(i,j) = \sqrt[q]{(|x_{i_1} - x_{j_1}|^q + |x_{i_2} - x_{j_2}|^q + \dots + |x_{i_p} - x_{j_p}|^q)}$$

anche detta **metrica L_q**

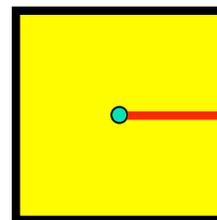
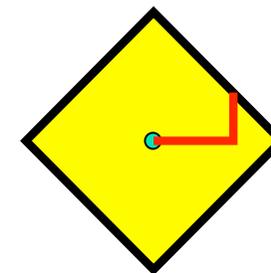
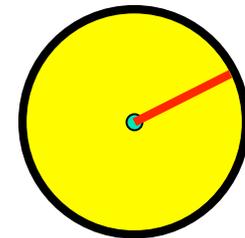
- $q = 2$, distanza euclidea

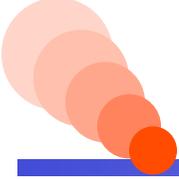
$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \dots + |x_{i_p} - x_{j_p}|^2)}$$

- $q = 1$, Manhattan distance

$$d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \dots + |x_{i_p} - x_{j_p}|$$

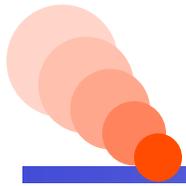
- $q = \infty$, massima differenza sulle coordinate





Proprietà delle distanze

- Una qualsiasi funzione tale che:
 - $d(i,j) \geq 0$
 - $d(i,i) = 0$
 - $d(i,j) = d(j,i)$
 - $d(i,j) \leq d(i,k) + d(k,j)$ per ogni k
(diseguaglianza triangolare)



Attributi binari

- Data una tabella di contingenza

		Oggetto j		
		1	0	<i>sum</i>
Oggetto i	1	a	b	$a+b$
	0	c	d	$c+d$
	<i>sum</i>	$a+c$	$b+d$	p

- conteggio dei **mismatching**

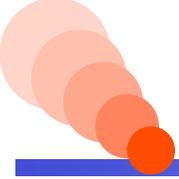
(se la semantica di 0/1 è simmetrica):

$$d(i, j) = \frac{b+c}{a+b+c+d}$$

- coefficiente di Jaccard

(se la semantica di 0/1 è asimmetrica):

$$d(i, j) = \frac{b+c}{a+b+c}$$

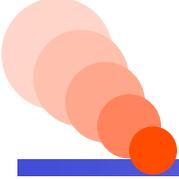


Attributi categorici

- Conteggio di mismatch
 - m : # matches, p : # attributi

$$d(i, j) = \frac{p - m}{p}$$

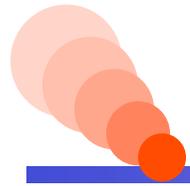
- Creare un attributo binario per ogni possibile valore dell'attributo (semantica simmetrica o asimmetrica)



Attributi ordinali

- E' rilevante l'ordine tra i valori, non il valore in sè

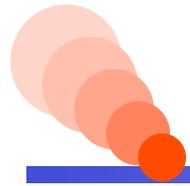
- Possono essere visti come numerici
 1. sostituendoli con il rank
 2. normalizzando nell'intervallo [0; 1]
 3. calcolando la dissimilarità attraverso una distanza



Attributi su scale ampie

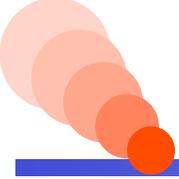
- In campo scientifico, economico, biologico, medico, ...

- Metodi:
 - valori numerici **INADEGUATO! (differenze assolute?)**
 - trasformazione logaritmica, poi considerati numerici
$$y_{if} = \log(x_{if})$$
 - valori ordinali - rank.



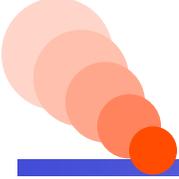
Attributi di tipo misto

- Si può ottenere la distanza attraverso una combinazione (pesata) delle distanze ottenute sui diversi tipi di attributi



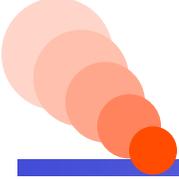
Cluster Analysis

- Introduzione
- Tipi di dato nella cluster analysis
- **Approcci**
- Metodi basati sul partizionamento
- Metodi gerarchici
- Metodi basati sulla densità
- Scoperta di outlier



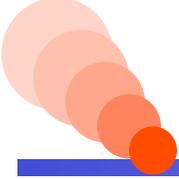
Approcci al clustering

- Partizionamento: costruiamo una partizione degli oggetti (o meglio, una famiglia di insiemi disgiunti) in base a misure di similarità
- Algoritmi gerarchici: attuiamo una decomposizione gerarchica, dall'alto o dal basso
- Algoritmi basati sulla densità: basato su soglie di densità in intorni degli oggetti
- Algoritmi grid-based: densità di regioni definite da un partizionamento dello spazio
- Algoritmi basati su modelli: si ipotizza un modello per ciascun cluster e si cercano insiemi che lo approssimano (p.es. riconoscimento di forme)



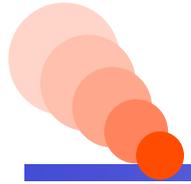
Cluster Analysis

- Introduzione
- Tipi di dato nella cluster analysis
- Metodi principali
- **Metodi basati sul partizionamento**
- Metodi gerarchici
- Metodi basati sulla densità
- Scoperta di outlier



Algoritmi di partizionamento

- Costruire una partizione di n oggetti in k cluster, ottimizzando qualche criterio (k fissato a priori)
 - ottimo globale: enumerazione di tutte le possibili partizioni
 - metodi euristici: *k-means* e *k-medoids*
 - *k-means* (MacQueen'67): il cluster è rappresentato dal suo centro
 - *k-medoids* o PAM (Partition Around Medoids) (Kaufman & Rousseeuw'87): il cluster è rappresentato dall'oggetto "mediano"

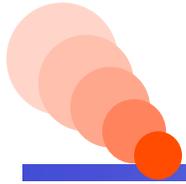


L'algoritmo k -means

Fissato il numero di cluster k :

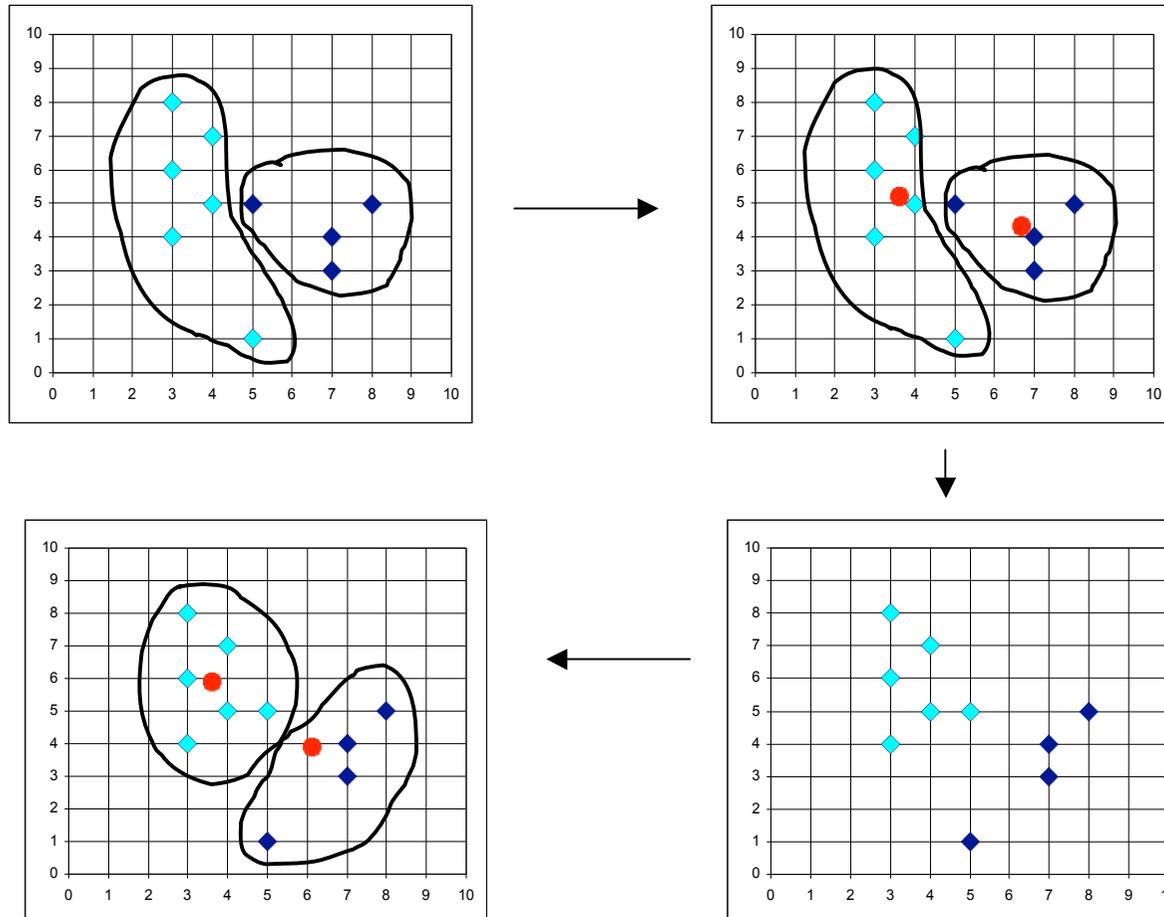
1. partiziona oggetti in k insiemi;
2. calcola i centroidi dei k insiemi (media di ciascuna coordinata);
3. assegna ogni oggetto al cluster con centroide più vicino;
4. ripeti dal punto 2, finché la partizione non cambia

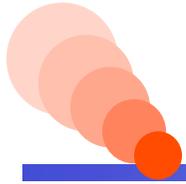




L'algoritmo k -means

$k=2$, partizione arbitraria





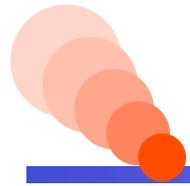
L'algoritmo k -means

- complessità: $O(tkn)$, dove
 - n = # oggetti,
 - k = # cluster,
 - t = # iterazioni (in genere $k, t \ll n$)
 - migliorabile con indicizzazione spaziale
- trova ottimo locale. Migliorabile con iterazioni ripetute, randomizzazione.

- usa il concetto di media. Non sempre applicabile
- numero di cluster predefinito
- clusterizza anche rumore e outliers
- solo cluster convessi

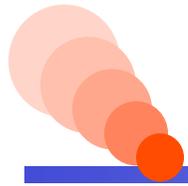
PRO

CONTRO



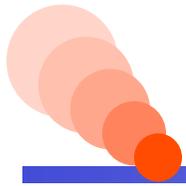
Varianti del *k*-means

- Selezione dei semi
- Calcolo delle distanze dai cluster (dissimilarità)
- Calcolo differenziale delle medie
- Dati categorici: *k*-modes (Huang'98)
 - usa la **moda** invece della **media** per il centro del cluster
 - cambiare distanza
 - *k*-prototype



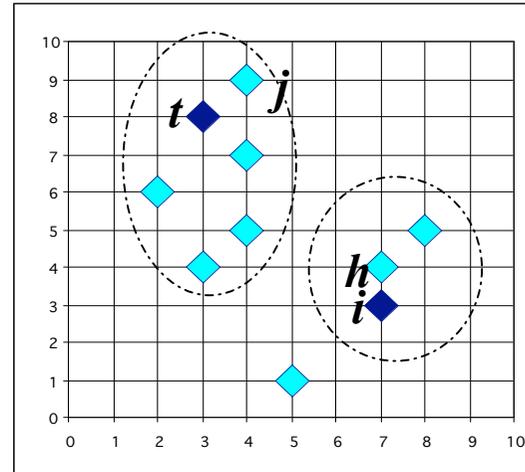
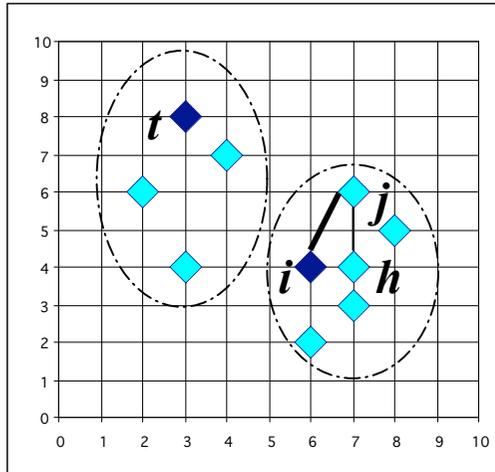
L'algoritmo *k*-medoids

- Cerca oggetti rappresentativi dei cluster
- *PAM* (Partitioning Around Medoids, 1987)
 - parte da un insieme iniziale di medoids
 - rimpiazza iterativamente un medoid se migliora la distanza totale del clustering
 - non scalabile
- *CLARA* (Kaufmann & Rousseeuw, 1990)
- *CLARANS* (Ng & Han, 1994): random sampling
- usando strutture ad indice: $O(n^2) \Rightarrow$ quasi $O(n)$ atteso



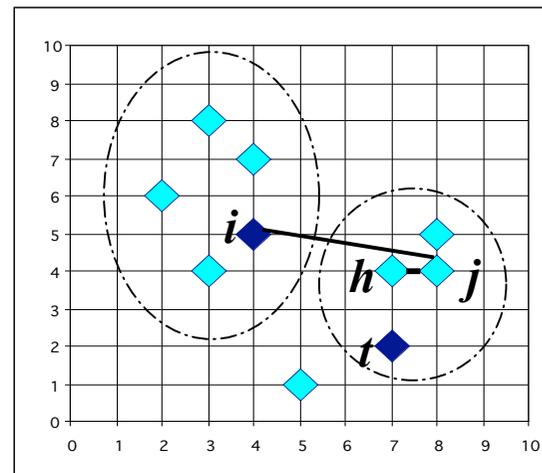
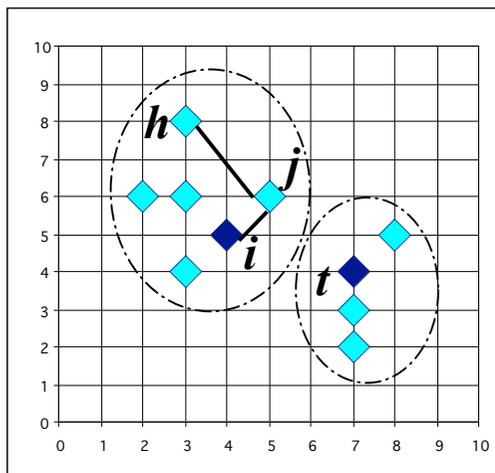
PAM Clustering: rimpiazzo medoid i con j ?

sfavorevole
rispetto a h

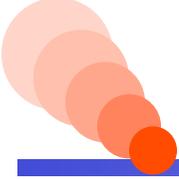


sfavorevole
rispetto a h

favorevole
rispetto a h

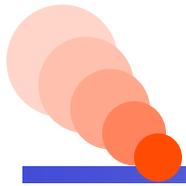


favorevole
rispetto a h



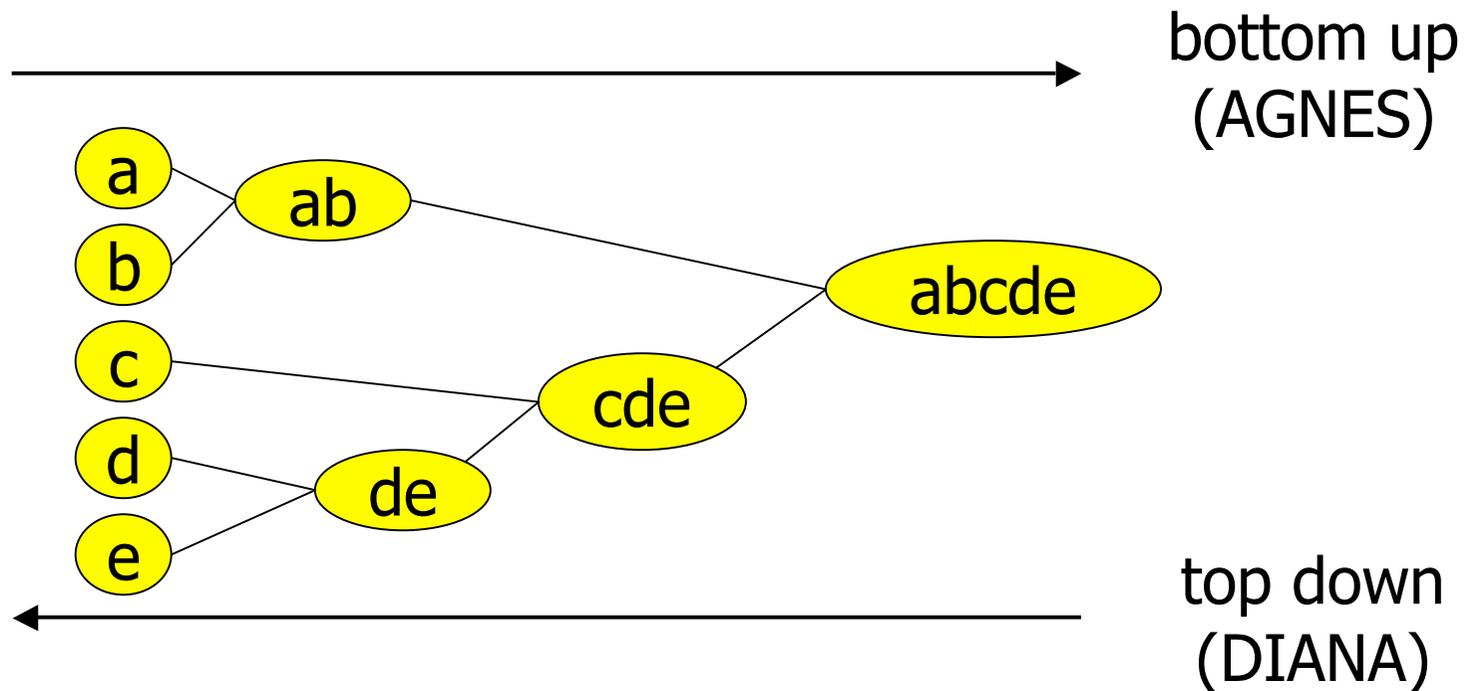
Cluster Analysis

- Introduzione
- Tipi di dato nella cluster analysis
- Approcci
- Metodi basati sul partizionamento
- **Metodi gerarchici**
- Metodi basati sulla densità
- Scoperta di outlier



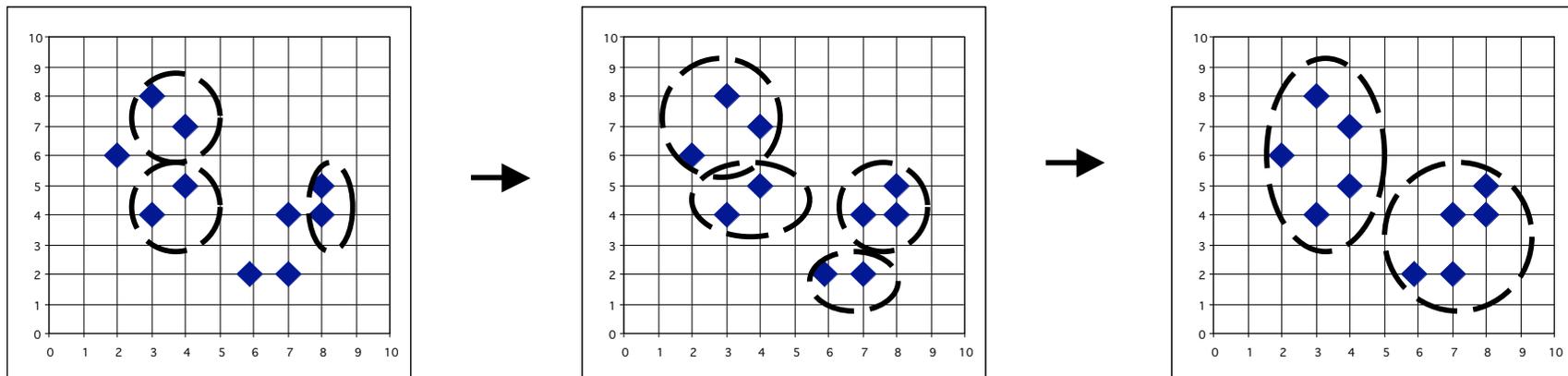
Clustering gerarchico

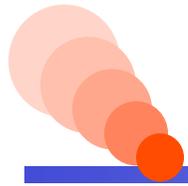
- Numero di cluster non fissato.
- Necessita di un criterio di arresto



AGNES (AGglomerative NESting)

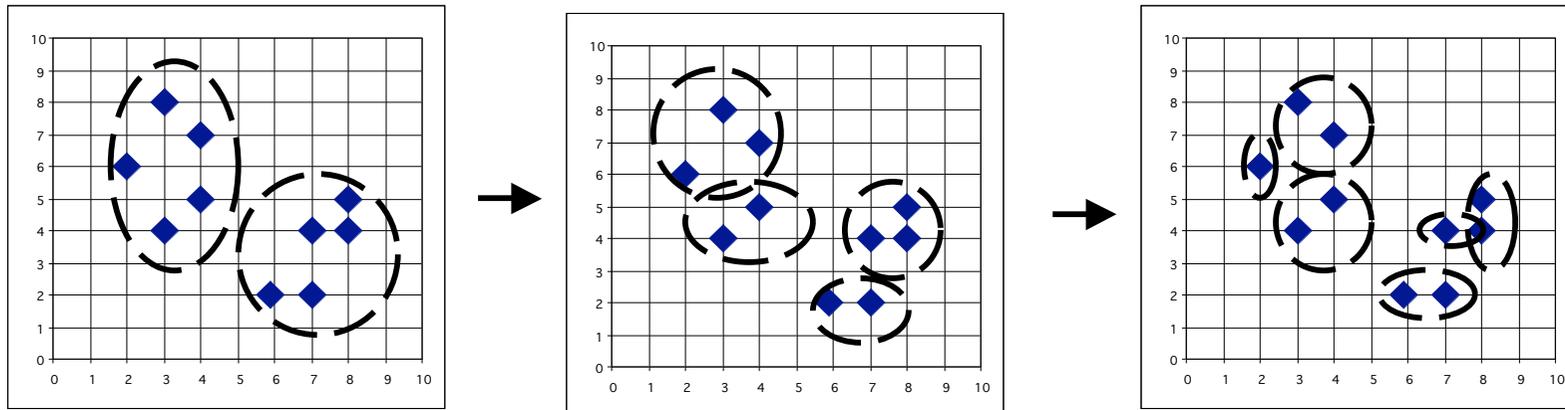
- [Kaufmann and Rousseeuw, 1990]
- disponibile in Splus
- fonde i due cluster che hanno minima dissimilarità
- criterio di arresto: soglia sulla dissimilarità di fusione
- numero di cluster variabile (dipende dalla istanza e dal criterio di arresto)

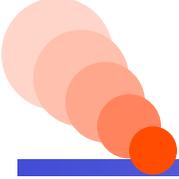




DIANA (DIvisive ANALysis)

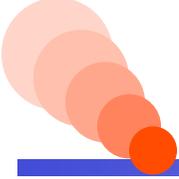
- [Kaufmann and Rousseeuw, 1990]
- disponibile in Splus
- processo inverso ad AGNES





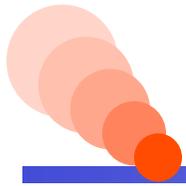
Altri metodi gerarchici

- AGNES e DIANA: complessità $O(n^2)$
- Metodi distance based con approccio “divide et impera”:
 - BIRCH (1996): usa CF-tree (Cluster Feature tree)
 - CURE (1998): sceglie punti rappresentativi del cluster, e li perturba verso il centro
 - CHAMELEON (1999): clustering gerarchico basato su k -nearest neighbor



BIRCH (1996)

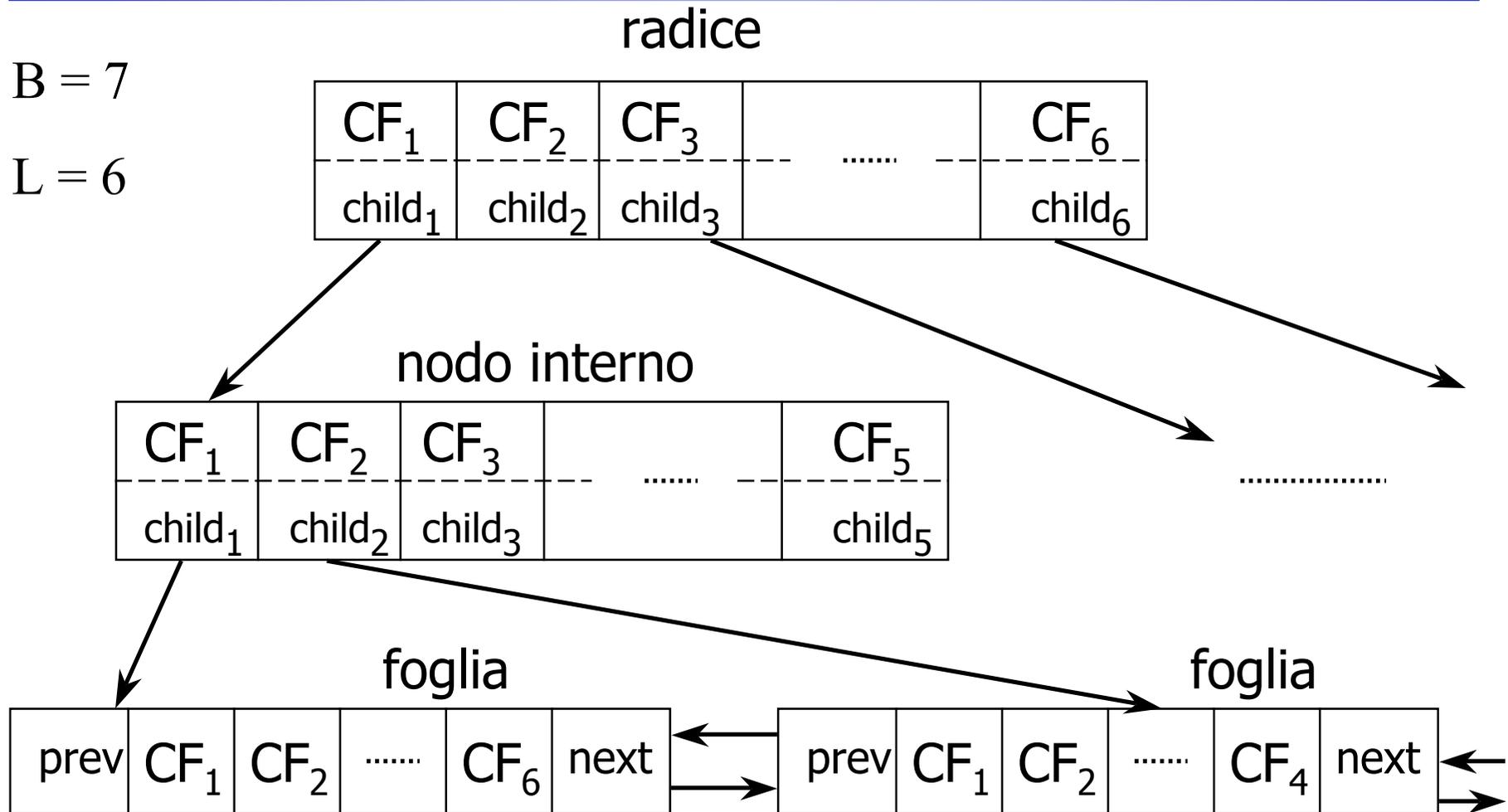
- BIRCH: Balanced Iterative Reducing and Clustering using Hierarchies, [Zhang, Ramakrishnan, Livny, SIGMOD'96]
- Decomposizione gerarchica basata su CF-tree (Clustering Feature). Memorizza momenti di sottoinsiemi.
 - fase 1: scansione lineare per costruire il CF-tree (leggermente superlineare)
 - fase 2: agglomera partendo dalle foglie del CF-tree con un algoritmo di clustering "locale"

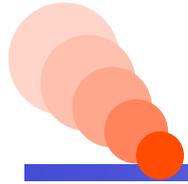


CF Tree

$B = 7$

$L = 6$

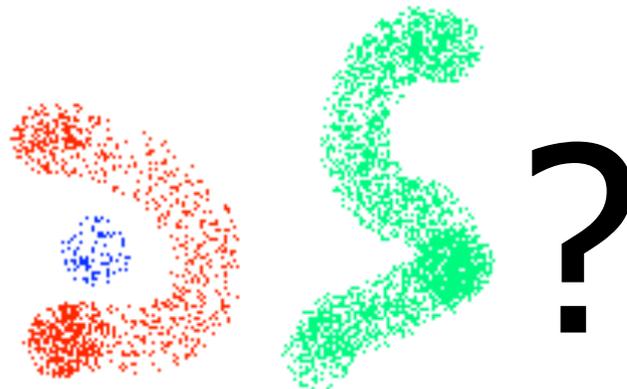




Problemi comuni

Il numero di cluster è quasi sempre predeterminato

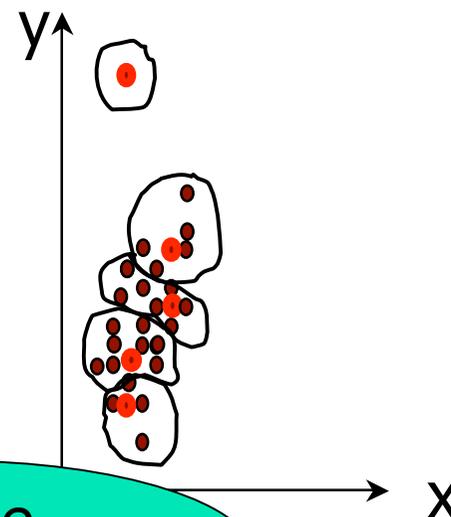
Un punto rappresentativo per ciascun cluster



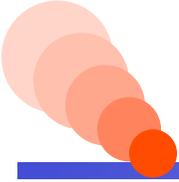
Cluster sferici, o almeno convessi

CURE (Clustering Using Representatives)

- Random sample s .
- Partiziona s in p sottoinsiemi bilanciati
- Crea cluster di q punti
- Elimina outliers
 - random sampling
 - cluster a crescita lenta
- Clusterizza clusters

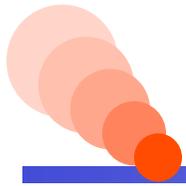


forme
arbitrarie

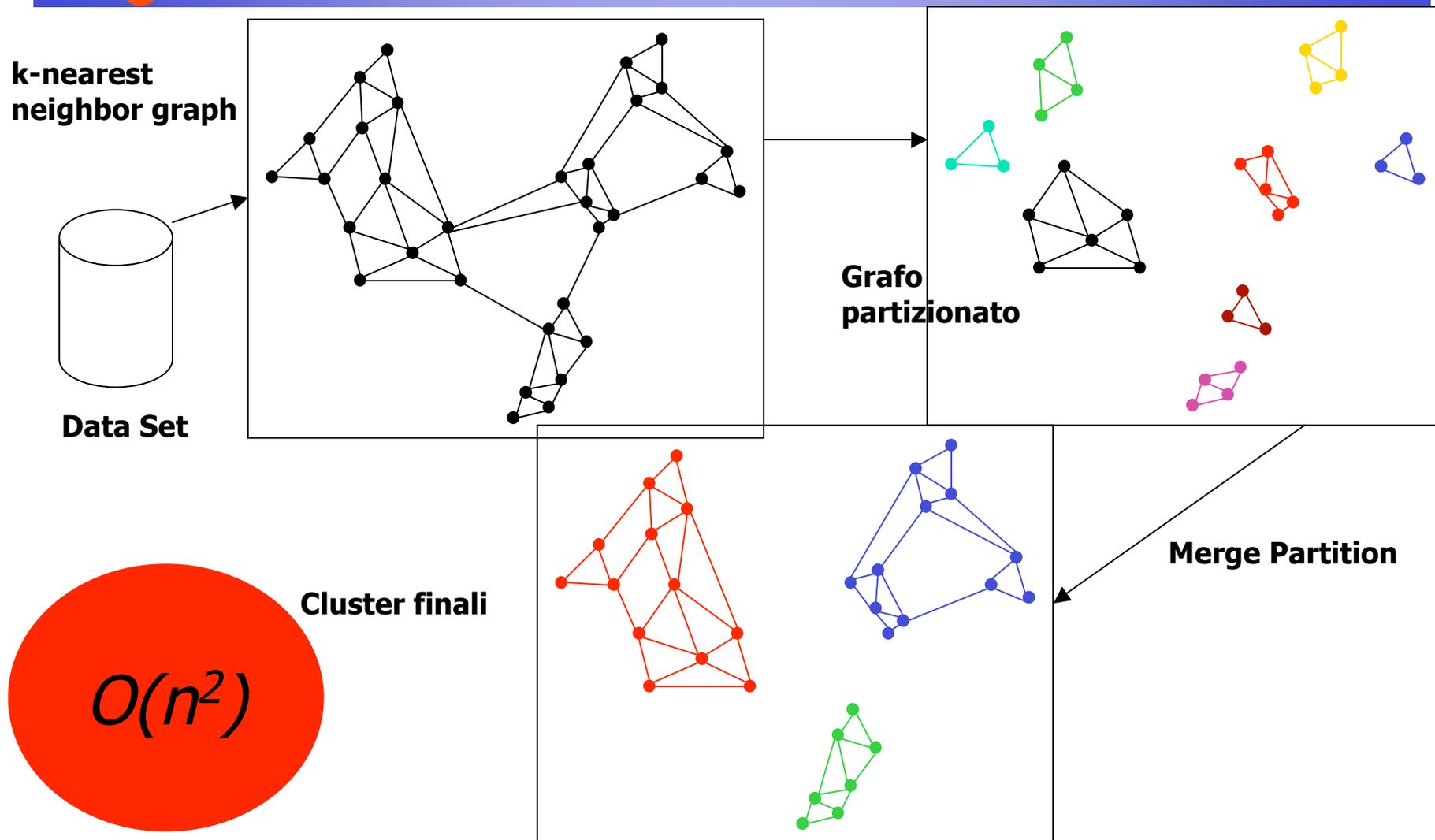


CHAMELEON (1)

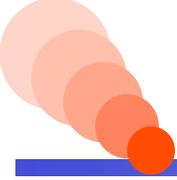
- [Karypis, Han, Kumar, 99]
- basato su un modello dinamico che misura dissimilarità
 - fonde due cluster solo se hanno *interconnettività* e *prossimità (proximity)* elevata rispetto alla interconnettività e prossimità interna ai cluster
- Algoritmo a due fasi:
 - 1. Partizionamento sul grafo: crea un gran numero di piccoli sub-cluster
 - 2. Algoritmo di agglomerazione gerarchica: fonde ripetutamente sub-cluster



CHAMELEON (2)

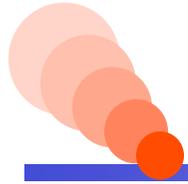


$O(n^2)$



Cluster Analysis

- Introduzione
- Tipi di dato nella cluster analysis
- Approcci
- Metodi basati sul partizionamento
- Metodi gerarchici
- **Metodi basati sulla densità**
- Scoperta di outlier



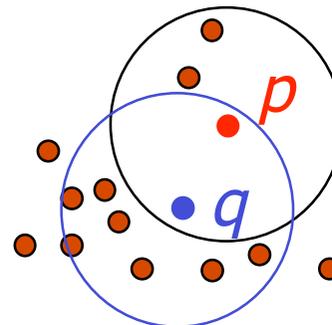
Metodi basati sulla densità

- Raggruppano in base alla densità locale di punti
- Caratteristiche:
 - cluster di forma arbitraria
 - isolano outliers
 - efficienti
 - parametri di input: soglie di densità

- DBSCAN: Ester, et al. (KDD'96)
- OPTICS: Ankerst, et al (SIGMOD'99).
- DENCLUE: Hinneburg & D. Keim (KDD'98)
- CLIQUE: Agrawal, et al. (SIGMOD'98)

Caratteristiche comuni ai metodi distance-based (1)

- Due parametri di base
 - *Eps*: raggio dell'intorno
 - *MinPts*: punti richiesti nell'intorno
- core point q : se ci sono almeno *MinPts* punti nell'intorno di q di raggio *Eps*
- Directly density-reachable: p è directly density-reachable da q se
 - 1) p è nell'intorno di q
 - 2) q è un core point



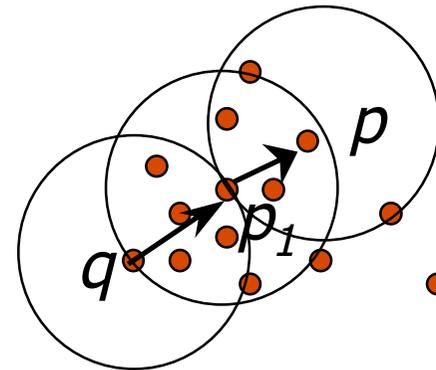
MinPts = 5

Eps = 1 cm

Caratteristiche comuni ai metodi distance-based (2)

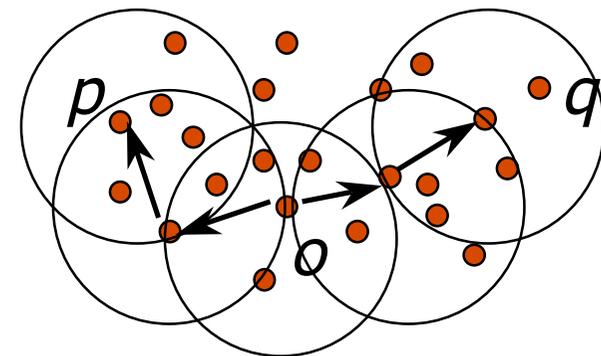
- density-reachability:

- chiusura transitiva della direct density-reachability



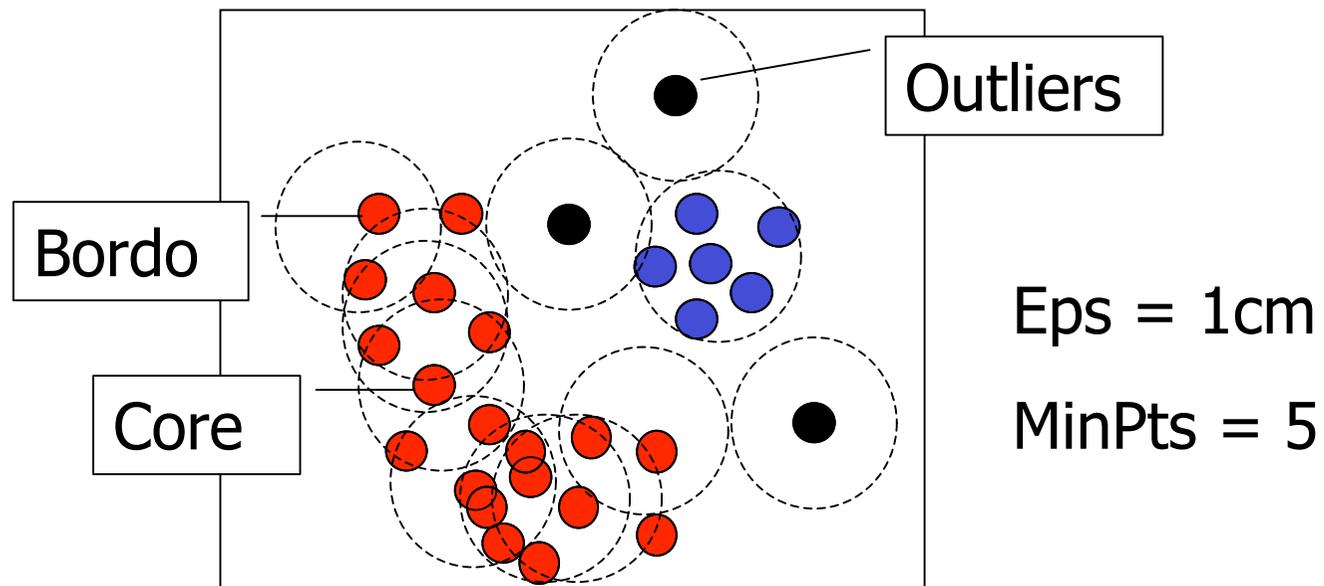
- *a* e *b* sono density-connected:

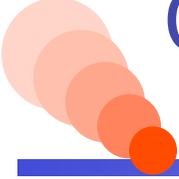
- se esiste un core point *c* dal quale sia *a* e *b* sono density-reachable



DBSCAN: Density Based Spatial Clustering of Applications with Noise

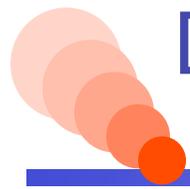
- Un cluster è un insieme massimale di punti density-connected
- Costruisce cluster di forma arbitraria
- Punti in zone non dense non appartengono a nessun cluster





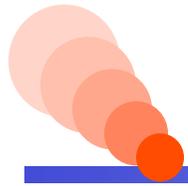
OPTICS: density-based parametrico

- OPTICS: Ordering Points To Identify the Clustering Structure
 - [Ankerst, Breunig, Kriegel, Sander, SIGMOD'99]
 - basato su DBSCAN, genera una sequenza di diversi clustering, al variare del parametro che definisce l'intorno
 - utile per applicazioni interattive, in cui l'utente può osservare il risultato al variare del parametro
- Stessa complessità di DBSCAN: $O(n \log n)$ con l'uso di strutture di accesso



DENCLUE: clustering con funzioni di densità

- DENSity-based CLUstEring [Hinneburg, Keim, KDD'98]
- Approccio
 - ciascun punto genera una funzione di influenza nello spazio circostante (p.es. gaussiana, o cilindrica)
 - si ha un cluster se la somma delle funzioni di influenza supera una soglia fissata
- Vantaggi
 - ottimo in presenza di rumore e outliers
 - applicabile anche in dimensioni elevate
 - molto efficiente (se applicato localmente)



Center-Defined and Arbitrary

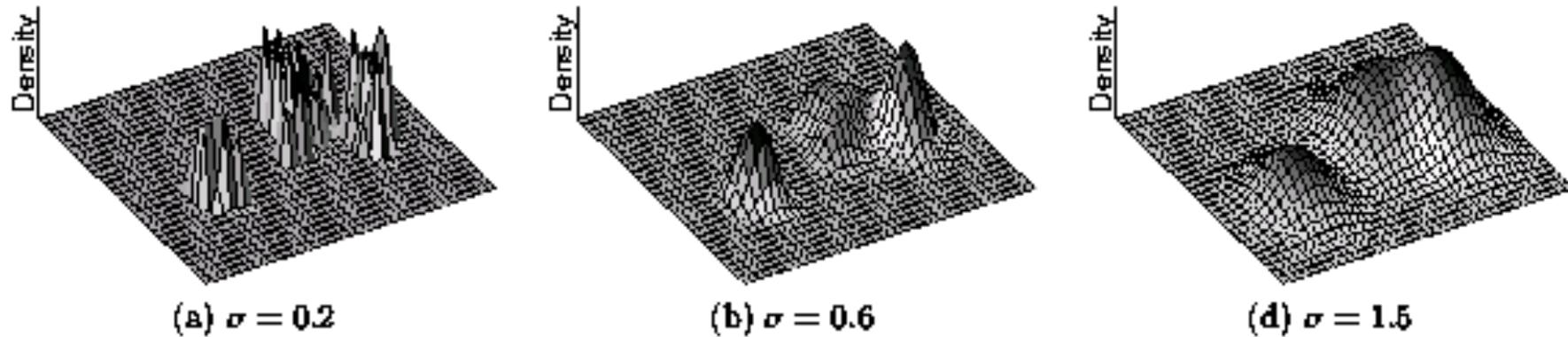


Figure 3: Example of Center-Defined Clusters for different σ

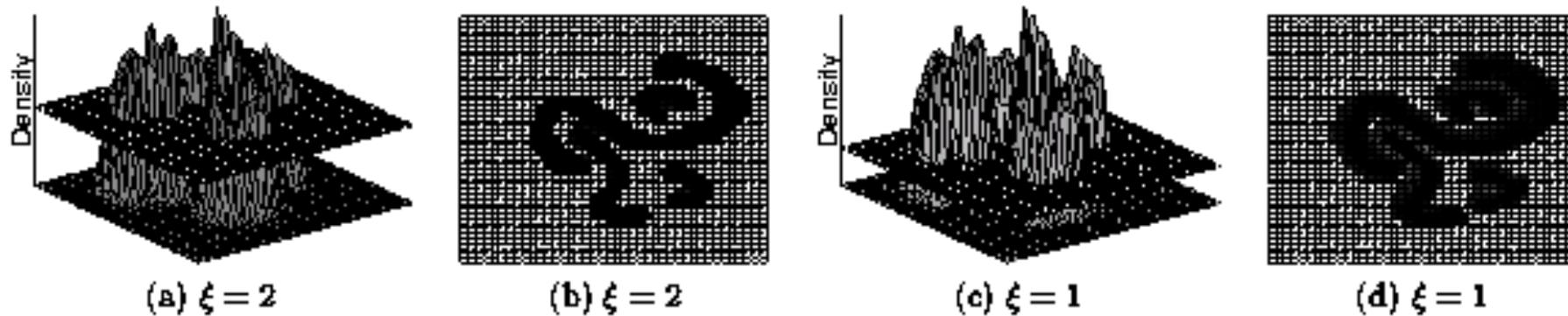
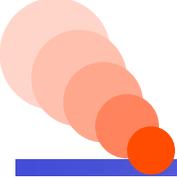
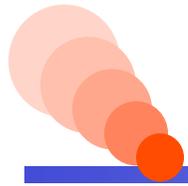


Figure 4: Example of Arbitrary-Shape Clusters for different ξ



Cluster Analysis

- Introduzione
- Tipi di dato nella cluster analysis
- Approcci
- Metodi basati sul partizionamento
- Metodi gerarchici
- Metodi basati sulla densità
- Scoperta di outlier

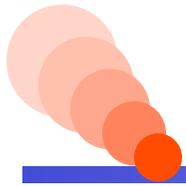


Outliers

- Outlier: elemento anomalo, non classificabile

- Problema:
 - trovare i k punti più anomali
 - eliminare tutti i punti anomali

- Applicazioni:
 - fraud detection
 - data cleaning
 - casi “clinici”



Clustering in presenza di vincoli

P.es. la presenza di collegamenti stradali e/o ostacoli rende inadeguata la distanza euclidea (è ancora una metrica!)

