Nuovi metodi di Cluster Analysis

Analisi dei gruppi e categorizzazione dei testi

La cosiddetta *categorizzazione* di documenti è uno dei più tipici compiti del Text Mining

Si tratta di risolvere il problema di assegnare una categoria a ciascun testo di una raccolta

Alle volte è, invece, necessario identificare soltanto una categoria di documenti (ad es. *spam* nelle *e-mail*, siti da evitare per bambini)

In genere la categoria è relativa al contenuto (*topic*) del documento o alla presenza di particolari riferimenti

Si tratta di un problema tipico dell'Information Retrieval, ma anche per obiettivi di analisi

Utilizzando il linguaggio della *computer science* si distingue in:

- ***classificazione supervisionata***

quando esiste una ***informazione esterna*** relativa alla ***corretta classificazione*** dei documenti, sia in termini di conoscenza delle categorie presenti, sia in termini di corrispondenza categorie - documenti

- ***classificazione non supervisionata***

quando non si dispone di informazione che non sia contenuta nel *corpus*

Le tecniche di cluster analysis affrontano il problema della classificazione non supervisionata

Nota - Si parla di una classificazione *semi-supervisionata*, quando si dispone dell'informazione relativa alla categoria per una parte dei documenti, generalmente utilizzata per la fase di apprendimento di regole di assegnazione

La *cluster analysis*

Le tecniche di c*luster analysis* (in italiano *analisi dei gruppi*) rappresentano uno dei temi principali dell'analisi dei dati, perché, da un lato, rispondono all'obiettivo classificatorio che viene generalmente posto alla base di un processo conoscitivo e, dall'altro, consentono di ottenere una sintesi dell'informazione disponibile

L'obiettivo della c*luster analysis* è, quindi, raggruppare i dati in gruppi (*cluster*) ***utili*** e/o ***interessanti***

L'idea di fondo è quella di catturare una ***struttura*** intrinsecamente presente nei dati

Nel caso di dati testuali, ad esempio, un tipico obiettivo di *cluster* è quello di identificare documenti collegati da un argomento comune (*topic*)

Richiami alle nozioni di base della *cluster analysis*

- Matrice dei dati e matrici di prossimità (distanza, dissimilarità...)
- Grafo di prossimità
- Misure e indici di similartà e di distanza (tipi di attributi e scale)
- Metrica Euclidea e varianti
- Classificazione gerarchica e non gerarchica

Questioni specifiche degli algoritmi di *clustering gerarchico*:

la distanza fra due *cluster* è calcolata secondo il tipo di legame:

a) *legame semplice* (distanza minima)

b) *legame completo* (distanza massima)

c) *legame medio* (distanza media)

il metodo di Ward

Richiami: l'algoritmo *k* medie

In presenza di basi di dati di grandi dimensioni, come accade nel *text mining*, si ricorre in genere ad ***algoritmi non gerarchici***, spesso derivati dall'antico, ma sempre valido, algoritmo *k-means*

L'algortmo *k-medie*, nella sua formulazione di base, si articola nei seguenti passi:

INPUT: matrice (*n*,*d*) individui × variabili; *k*, numero di gruppi da identificare

1. *si selezionano casualmente* k *individui, come centroidi iniziali*
2. *si calcola la distanza di ciascun individuo dai* k *centroidi e lo si assegna al centroide più vicino*
3. *si ricalcola il centroide di ciascun gruppo di individui*
4. *si ripetono i passi 2 e 3 fino a che i centroidi non cambiano (o cambiano di poco)*

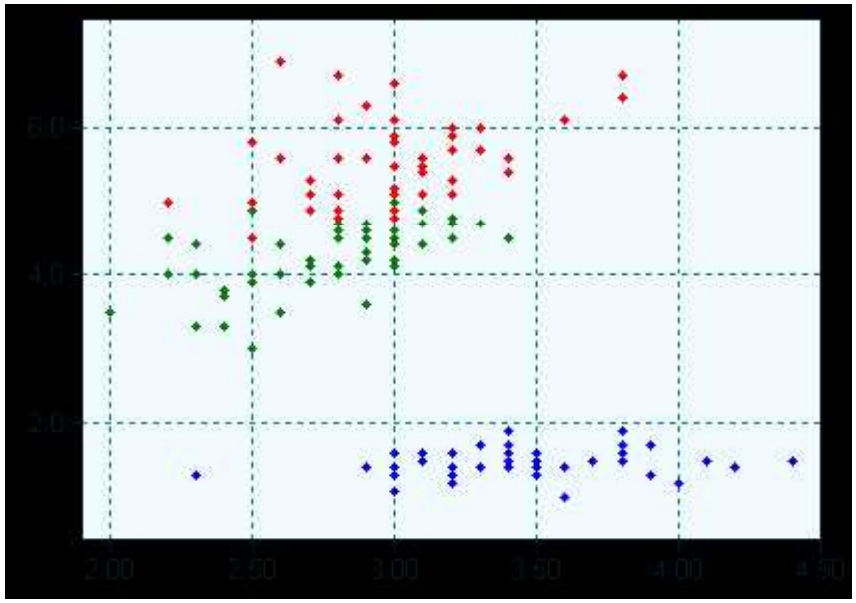
OUTPUT: *k* gruppi distinti

N.B. di base, per distanza intendiamo la distanza Euclide, ma l'algoritmo presenta numerosissime varianti, nelle quali, alle volte, vengono utilizzate distanze differenti

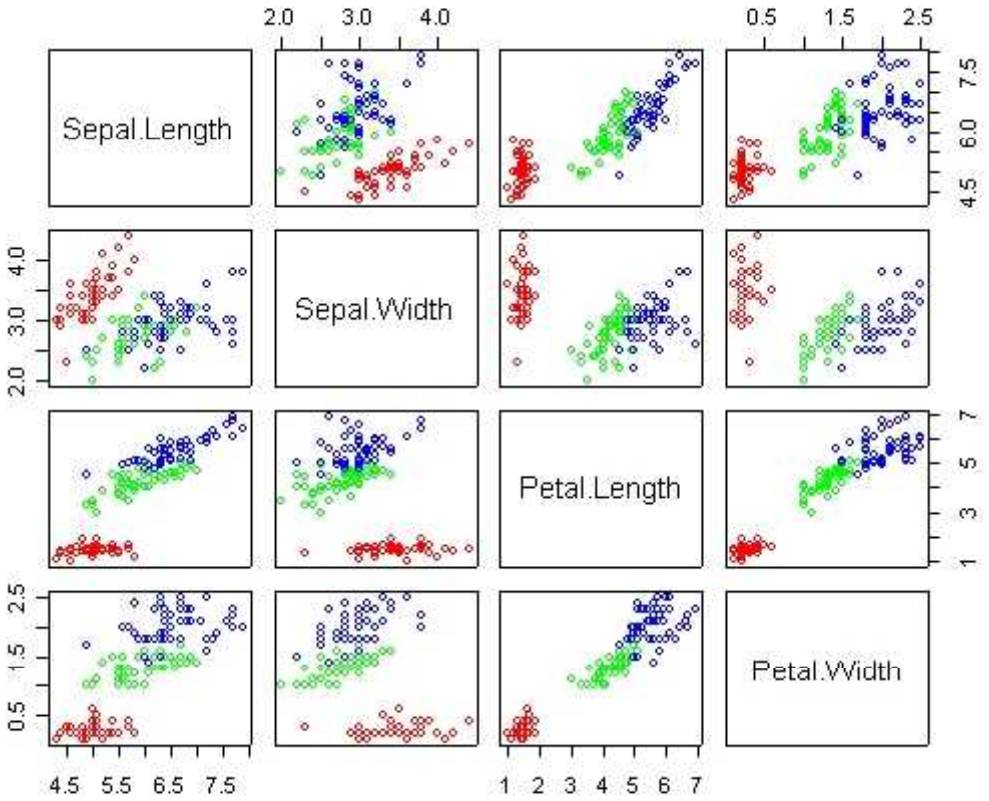
Richiami e definizioni

L'obiettivo dell'*analisi dei gruppi* è quello di mettere insieme gli oggetti studiati, sulla base delle loro caratteristiche o delle relazioni fra loro esistenti
Il risultato che si vuole ottenere è entificare **gruppi** quanto più possibile *omogenei* al loro interno ed *eterogenei* fra loro

Nella realtà, non sempre ci troviamo di fronte a situazioni in cui i gruppi sono nettamente costituiti, ma le tecniche di *cluster* generalmente (ad eccezione di quelle basate sulla logica *fuzzy*/sfocata) producono come risultato dei gruppi nitidi/*crisp*, privi di sovrapposizioni

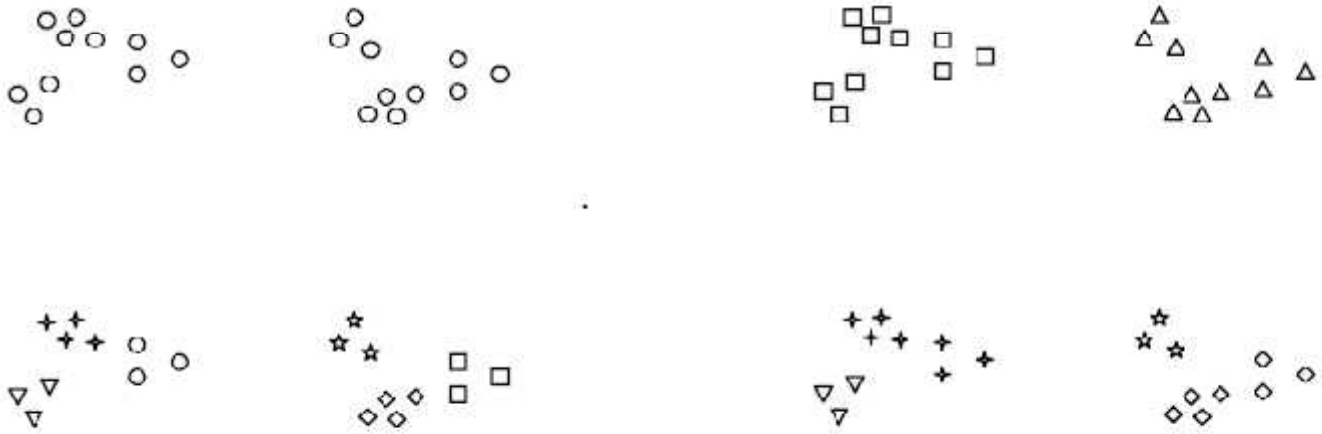


In uno degli esempi più famosi della statistica, gli "iris di Fisher", in assenza di informazioni esterne, saremmo probabilmente portati, da una lettura del piano che incrocia due descrittori (larghezza di petali e sepali) ad immaginare, ad esempio, la presenza di due classi



Scatter plot matrix degli iris di Fisher

Esistono i gruppi? E se sì quanti?



Due? Quattro? Sei? ???

Nota bene – E' la stessa nozione di gruppo ad essere imprecisa e strettamente legata alla natura dei dati ed agli obiettivi conoscitivi ed operativi che ci poniamo

Cluster Analysis per dati altamente dimensionali

Si parla in genere di

dati altamente dimensionali

quando la rappresentazione dei dati su cui vogliamo lavorare raggiunge centinaia se non migliaia di dimensioni

In altri termini, quando gli oggetti di cui vogliamo ottenere una classificazione sono caratterizzati da un numero estremamente elevato di descrittori

Nel linguaggio degli statistici, potremmo parlare di un numero elevatissimo di variabili, ma in questo contesto, come vedremo, il confine fra individui e variabili sembra sfumare ...

Dati altamente dimensionali

Gli ambiti in cui si trova più frequentemente questo tipo di problema sono l'analisi dell'espressione genica ed il Text Mining

Gli sviluppi tecnologici consentono, infatti, di descrivere il DNA in *micro-array* costituiti da numerosissime misurazioni, mentre in una codifica *bag-of-words* i vettori che descrivono i documenti hanno dimensione pari all'ampiezza del vocabolario

Stranamente due contesti apparentemente così distanti si trovano a sviluppare soluzioni metodologiche e algoritmiche sempre più spesso comuni, con effetti estremamente interessanti, anche ai fini di un'adeguata valutazione della qualità delle soluzioni proposte

La MALEDIZIONE della MULTIDIMENSIONALITA':
gli spazi multidimensionali sono vuoti!!!

Intuitivamente, quante più caratteristiche conosciamo di un individuo, tanto più ci appare diverso dagli altri individui appartenenti al collettivo oggetto di studio

Si parla di "curse of dimensionality" quando occorre fare analisi di dati descritti da un numero rilevante di variabili:
molte dimensioni sono impossibili da visualizzare e difficili da immaginare
In generale, si può dire che I problemi che occorre affrontare con dati altamente dimensionali sono collegati alla "sparsità" degli spazi in cui gli oggetti sono rappresentati

Immaginiamo di avere 100 punti distribuiti casualmente in un intervallo [0, 1] e di visualizzarli su una retta, suddivisa in 10 segmenti di lunghezza 0,1 (rappresentazione a 1 dimensione)

Ora, immaginiamo di rappresentarli in una tabella con 10 righe e 10 colonne (rappresentazione bidimensionale), ovvero su una griglia composta da quadratini di lato 0,1

Avremo così 100 celle e probabilmente alcune saranno vuote
Se adesso analogamente consideriamo una terza dimensione e 1000 celle, I nostri 100 punti cominceranno ad essere

“dispersi nello spazio”

.....

Il problema della DISTANZA

La stessa nozione di distanza diventa meno precisa al crescere del numero delle dimensioni e l'idea stessa di punti più o meno vicini perde di significato da un punto di vista dell'identificazione di elementi che siano simili ai fini dell'obiettivo conoscitivo che stiamo perseguendo

Del resto le tecniche di *clustering* dipendono in genere criticamente dalla misura di distanza (o di *dis-similarità*) su cui sono basate al fine di identificare la presenza ed il numero di gruppi interessanti

In situazioni altamente dimensionali, è stato dimostrato (Beyer *et al.*, 1999) come, per alcune distribuzioni, la differenza fra la distanza di un punto dal punto più lontano e quella dal punto più vicino tenda a 0 al crescere del numero delle dimensioni considerate

$$\lim_{d \to \infty} \frac{MaxDist - MinDist}{MinDist} = 0$$

Si dice infatti che “in spazi altamente dimensionali, le distanze tra punti diventano relativamente uniformi” e la nozione di “nearest neighbor” perde di significato

Richiamiamo la distanza di Minkowski p_{ij} fra 2 punti x_i e x_j

$$p_{ij} = \left(\sum_{k=1}^{d} \left| x_{ik} - x_{jk} \right|^{r} \right)^{1/r}$$

dove *d* è la dimensione e *r* il parametro che definisce il tipo di distanza. Per *r* =1 si ha la cosiddetta distanza di Manhattan, per *r*=2 la consueta distanza euclidea …

Per r >=3 è stato dimostrato (Hinneburg et al., 2000) che

$$\lim_{d \to \infty} MaxDist - MinDist = 0$$

Scelta delle variabili rilevanti ai fini del problema di classificazione

local feature relevance

Uno dei primi modi per affrontare il problema dell'eccessiva dimensionalità del fenomeno da sottoporre ad una procedura di *clustering* è quello di ricorre ad informazioni *a priori* al fine di ridurre il numero di variabili da considerare

Più in generale, in tutte le procedure di *mining*, è importante procedere con una fase preliminare che consenta la cosiddetta *feature selection*, al fine di rimuovere informazione "irrilevante" o "ridondante", ad esempio perché con una modesta variabilità, oppure con una forte correlazione con altre variabili considerate

Si tratta di una fase molto delicata e complessa

Principali strumenti per il *clustering* di dati altamente dimensionali

Riduzione della dimensione

Il problema viene affrontato di solito in termini di:

- ***feature extraction***

quando si costruisce un numero ridotto di variabili “latenti” che contengono una parte rilevante dell'informazione disponibile

Si ricorre, di solito, in qualche modo alla decomposizione in valori singolari di Eckart&Young (1936), che come è noto è alla base della maggior parte dei metodi statistici di analisi multivariata, come ad esempio l’Analisi delle Corrispodenze, nonché del Latent Semantic Indexing e, quindi, della Latent semantic Analysis

Si tratta di costruire un numero ridotto di variabili “latenti” che contengono una parte rilevante dell'informazione disponibile

oppure di:

- ***feature selection***

quando si seleziona, all'interno delle variabili “manifeste”, quelle più efficaci ai fini dell'identificazione di *cluster*
Nell’ambito del *text mining*, molte operazioni attuate in fase di pre-trattamento, come l’introduzione di una *stop list*, o le scelte di lemmatizzazione, oppure *stemming*, o, ad esempio, la scelta di lavorare con le sole POS lessicali, sono forme di *feature selection*

I metodi di riduzione della dimensionalità sono in genere molto vantaggiosi in termini di costi computazionali, ma rischiano di distruggere informazione rilevante ai fini della identificazione dei gruppi

E' noto, infatti, che i gruppi sono di solito nascosti in diversi sottospazi dello spazio a piene dimensioni ed un approccio generale rischia di essere pericoloso

Modelli parsimoniosi

Un altro strumento consiste nel ricorso a modelli che descrivano i dati con un numero esiguo di parametri da stimare

L'idea di base è che i dati possano essere modellati secondo distribuzioni gaussiane differenti nelle diverse classi

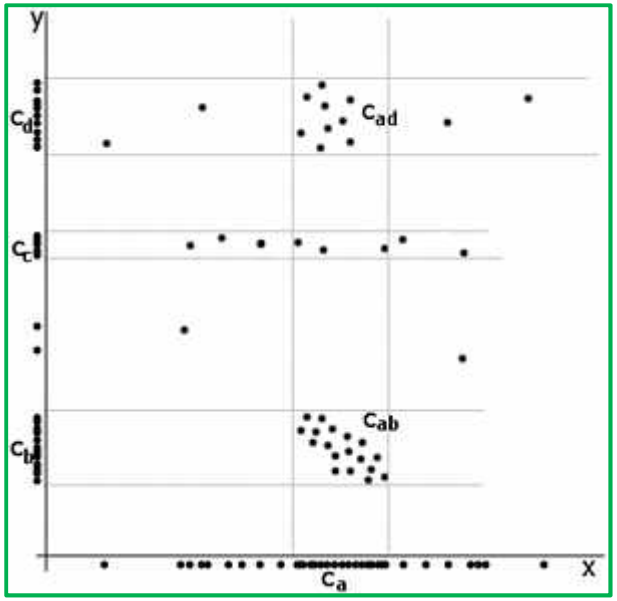
Se ogni classe è rappresentata da una Gaussiana (ϕ è una normale *p*-variata di parametri $\theta_i = \{\mu_i, \Sigma_i\}$

$$f(x, \theta) = \sum_{i=1}^{k} \pi_i \phi(x, \theta_i),$$

con $_i$ parametro di mistura

Introducendo alcune semplificazioni (l'estrema consiste nell'assumere una struttura di covarianza sferica per tutti i gruppi), si procede con una sorta di estensione dell'algoritmo *k-medie*, all'interno di un approccio EM

Clustering in sottospazi



L'immagine (presa da Wikipedia) mostra come in uno spazio a due dimensioni si identificano i gruppi $\boldsymbol{c_{ad}}$ *e* $\boldsymbol{c_{ab}}$
Se si guarda al sottospazio x in ascissa vedremo soltanto un grande gruppo $\boldsymbol{c_a}$
Se si guarda al sottospazio y in ordinata vedremo 3 gruppi $\boldsymbol{c_b}$ $\boldsymbol{c_c}$ $\boldsymbol{c_d}$

Il problema è che i punti appartengono a gruppi differenti a seconda della dimensione del sottospazio considerato

Il *clustering in sottospazi* è un particolare caso di *feature selection*

In genere gli algoritmi basati sul *clustering in sottospazi* fanno ricorso a qualche ragionamento di tipo euristico, quale ad esempio porre la cosidetta proprietà *downward closure*, che può essere semplificata come “se uno spazio *k* dimensionale è denso, anche un suo sottospazio sarà denso”

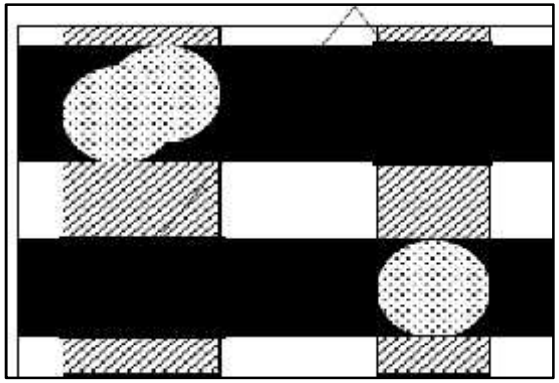
Prendiamo ad esempio l'algoritmo CLIQUE di Agrawal *et al. (1999)*

L'algoritmo CLIQUE di Agrawal, Gehrke, Gunopulos , Raghavan, 1998

L'idea di base è che una regione che è densa in un particolare sottospazio deve produrre regioni dense quando viene proiettata in un sottospazio di dimensioni ridotte

Per esempio, esaminando la figura vediamo degli addensamenti unidimensionali (in orizzontale ed in verticale)

Le strisce scure orizzontali e quelle tratteggiate in verticale suggeriscono la presenza, nella loro intersezione di *cluster* bidimensionali



Così, CLIQUE parte individuando intervalli densi sulla retta e successivamente passa ad esaminare il piano alla ricerca di zone dense bidimensionali

La procedura si itera a sottospazi di dimensione maggiore e lo fa in maniera più efficiente rispetto ad una ricerca rivolta alla ricerca di aree dense in tutti i possibili sottoinsiemi dimensionali

Clustering proiettato

L'idea di base è che si vuole assegnare ogni punto ad un gruppo, ma i gruppi possono esistere in sottospazi diversi

Esistono due approcci principali:

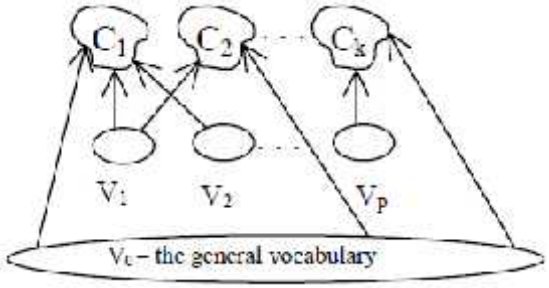
- si assume che i centroidi delle classi si trovino in un sottospazio comune, ignoto (*projection pursuit)*
- si assume che ogni classe si trovi in un suo sottospazio specifico (analisi in componenti principali)

Es. analisi fattoriale tipologica di Diday, basata su un algoritmo simile al *k*-medie

A questa classe appartiene anche l'algoritmo **PROCLUS** di Agrawal *et al. (2000)*, basato su una variante dell'algoritmo *k-medie*, facendo riferimento però, come centro, piuttosto che ad un individuo "medio", al cosiddetto "medoide", che è un individuo centrale, effettivamente osservato

In questo caso la distanza opportuna è quella di Manhattan (vista precedentemente)

Clustering basato su “concetti”



Un modo diverso per affrontare il problema consiste non nel cercare elementi che condividano attirbuti, ma piuttosto che condividano concetti, o meglio si tratta di trovare i gruppi in spazi di concetti e non in tradizionali spazi vettoriali

Un **concetto** può essere visto come un insieme di attributi

L'importanza di riferirsi a concetti è che gli oggetti possono essere concepiti come realizzazioni di concetti, in un'ottica di tipo probabilistico

Trattando di identificazione di gruppi di documenti, questi ultimi possono essere concepiti come composti da parole che appartengono ad uno o più concetti, intesi come insiemi di parole, o addirittura vocabolari, con la probabilità di ciascuna parola determinata da un modello statistico sottostante

La conseguenza è che si modifichi il modo di concepire la distanza fra documenti

In questo senso esistono numerose proposte che fanno riferimento a metodi statistici proposti per strutture di dati gerarchiche, come nei modelli *multilevel*, anche all'interno di impostazioni di tipo *bayesiano*

A questo approccio appartengono i cosiddetti *topic models,* che si basano su informazione esterna relativa agli argomenti di cui parlano i documento

Alcune considerazioni sul *High Dimensional Clustering*

Di fronte a situazioni complicate, come quelle che generalmente vengono affrontate quando si cercano gruppi in spazi altamente dimensionali e, ancor più quando il problema riguarda descrittori non numerici, come nel caso del *clustering* di documenti, molte categorie classiche dell'analisi dei dati, anche multidimensionali, vengono ad essere sfumate

Si parla, ad esempio, per i metodi basati su riduzioni dimensionali di algoritmi di "**soft clustering**", quando nel calcolo delle distanze gli attributi vengono ad essere ponderati differentemente, ma mai attribuendo un peso = 0, ovvero mai eliminando variabili (come nel caso della *feature selection*)

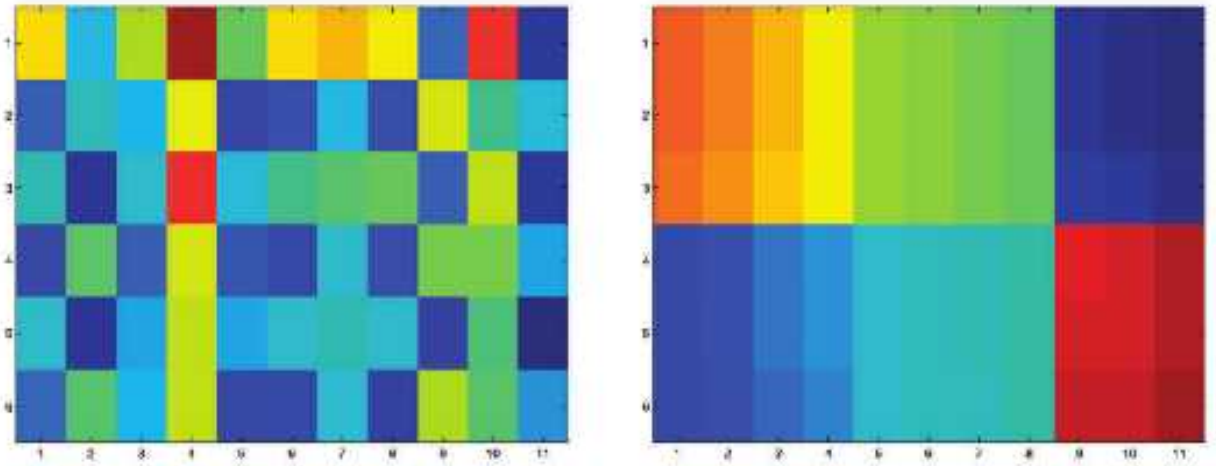
Ancora, si parla di approcci **ibridi** quando l'obiettivo non è di trovare una regola di assegnazione unica per ogni punto, consentendo sovrapposizioni di classi, ma anche accettando che alcuni oggetti non appartengano a nessun gruppo, contravvenendo alla classica definizione di partizione, per cui ciascun elemento deve appartenere ad una ed ad una sola classe (questa è una critica che viene mossa a CLIQUE)

Coclustering

Data:

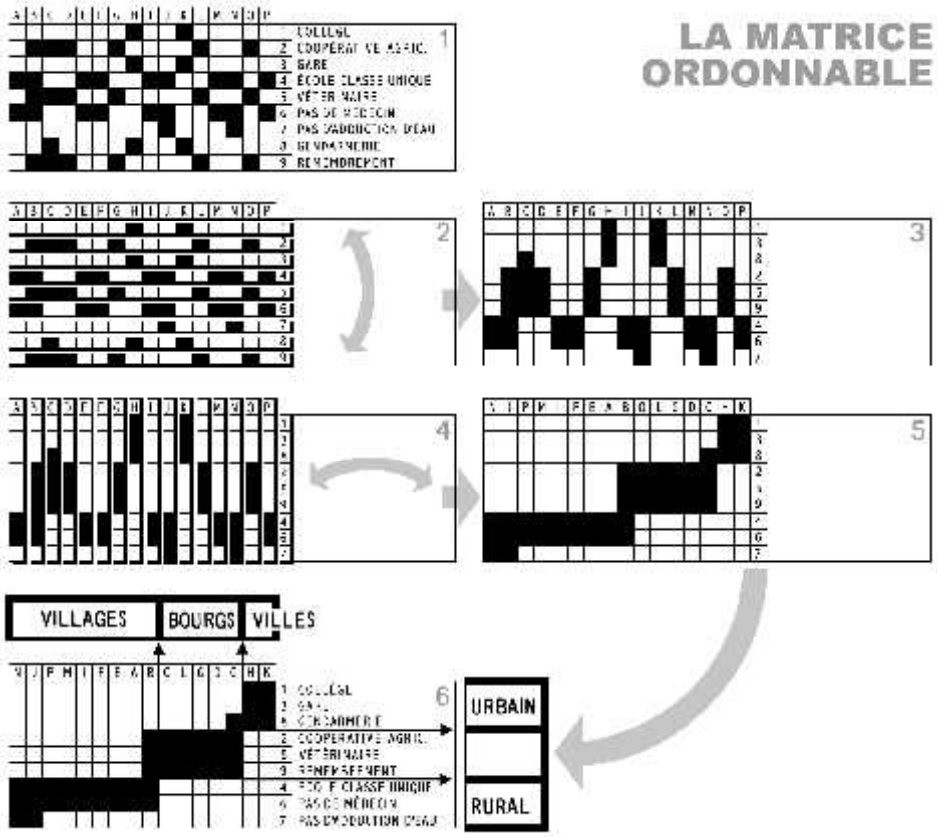
- Una partizione (I_1,..., I_r, ,.. I_R) di righe (R)
- Una partizione (J_1, . . ., J_c, . . ., J_C) di colonne (C)
- Una partizione *RxC* ottenuta incrociando la partizione di riga e la partizione di colonna

Righe e colonne della matrice dei dati sono partizionate in modo che i gruppi di riga e quelli di colonna siano vicini



Two-Way Clustering ⊙ *Co-Clustering* ⊙ *Simultaneous Clustering* ⊙ *Subspace Clustering* ⊙ *Bi-Dimensional Clustering* ⊙ *Bi-Clustering*

L'idea originale è antica e può farsi risalire ai grafici del geografo Jacques Bertin, inventore della semiologia grafica



Coclustering

«metodi di *cluster* in cui righe e colonne della matrice dei dati sono raggruppate contemporaneamente»

Nell'ambito della metodologia statistica, l'idea si fa risalire ad Hartigan (1972), che proponeva di identificare all'interno di una procedura alternata, gruppi di individui e gruppi di variabili, identificando gruppi di oggetti simili che possiedono gruppi di caratteristiche simili

Si tratta di un algoritmo a passi alterni, in cui si identificano successivamente gruppi di individui e gruppi di variabili, ossia, si raggrupano successivamente ed indipendentemente le righe e le colonne della matrice dei dati

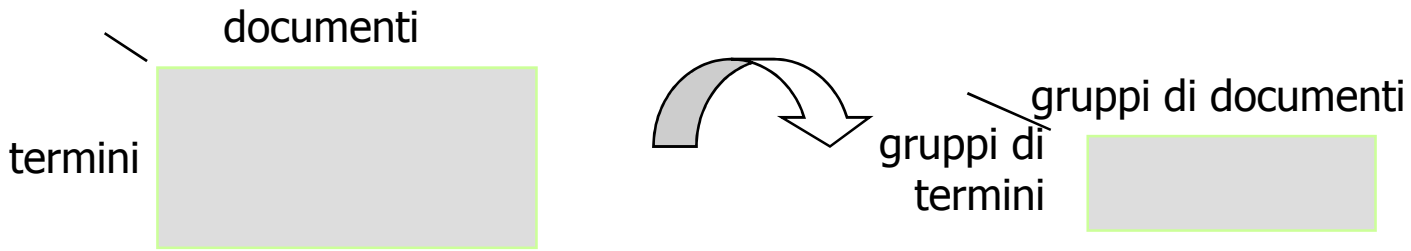
Oggi, invece, si preferiscono approcci simultanei, basati sulla ottimizzazione di una qualche funzione obiettivo comune

Sono moltissimi gli algoritmi proposti per ottenere il *coclustering* di grandi matrici sparse, così come spesso accade per tabelle lessicali, del tipo termini-per-documenti

In genere, le tabelle lessicali sono considerate una distribuzione di probabilità empirica congiunta per due variabili casuali discrete, vocabolario/collezione-di-documenti (cioè una tabella di contingenza)

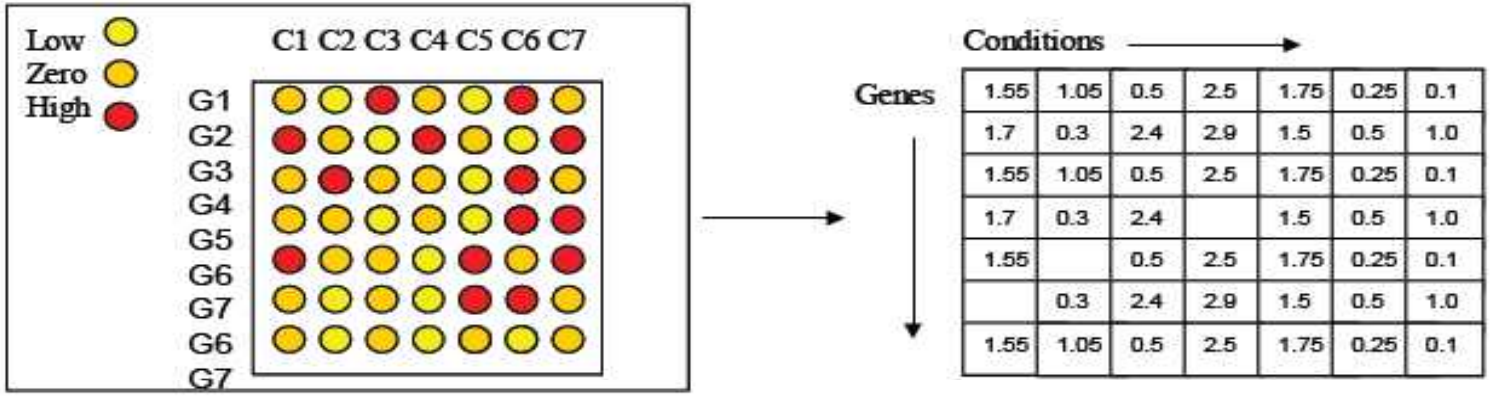
Volendo classificare I documenti, il cluster simultaneo di termini e documenti è motivato dall'assunzione che

gruppi di parole introducono il "contesto" nella procedura di identificazione di gruppi di documenti



Strutture di dati simili

Microarray: dati dell'espressione genica



1.55	1.05	0.5	2.5	1.75	0.25	0.1
1.7	0.3	2.4	2.9	1.5	0.5	1.0
1.55	1.05	0.5	2.5	1.75	0.25	0.1
1.7	0.3	2.4		1.5	0.5	1.0
1.55		0.5	2.5	1.75	0.25	0.1
	0.3	2.4	2.9	1.5	0.5	1.0
1.55	1.05	0.5	2.5	1.75	0.25	0.1

Si noti che anche questi dati subiscono, così come i dati testuali una massiccia fase di pretrattamento al fine di essere trasformati in codici numerici

Simile Struttura dei dati ⇨ Simili obiettivi

- **matrici sparse, dati altamente dimensionali**
- **#righe (termini/geni) >> #colonne (documenti/profili)**
- **Ha senso raggruppare sulle due vie**

In genetica i **geni (righe)** sono raggruppati rispetto al loro simile **profilo genetico (colonne)**	Nel TM i **documenti (colonne)** sono raggruppati rispetto a simili **profili lessicali (righe)**

Una rapida panoramica di algortmi di biclustering in ambito bio-medico

(da Jesús S. Aguilar-Ruiz)

Method	Publish	Allow overlap?	Complexity	Testing Data
Cheng & Church	ISMB 2000	Yes (rare in reality)	O(MN) or O (MlogN)	Yeast (2884×17), lymphoma (4026×96)
Getz et al. (CTWC)	PNAS 2000	Yes	Exponential	Leukemia (1753×72), colon cancer (2000×62)
Lazzeroni & Owen (Plaid Models)	Bioinformatics 2000	Yes	Polynomial	Food (961×6), forex (276×18), yeast (2467×79)
Ben-Dor et al. (OPSM)	RECOMB 2002	Yes	$O(NM^3l)$	Breast tumor (3226×22)
Tanay et al. (SAMBA)	Bioinformatics 2002	Yes	$O((N2^{d+1})^{\log_{(r+1)/r(rd)}})$	Lymphoma (4026×96), yeast (6200×515)
Yang et al. (FLOC)	BIBE 2003	Yes	$O((N+M)^2kp)$	Yeast (2884×17)
Kluger et al. (Spectral)	Genome Res. 2003	No	Polynomial	Lymphoma (1 rel., 1 abs.), leukemia, breast cell line, CNS embryonal tumor

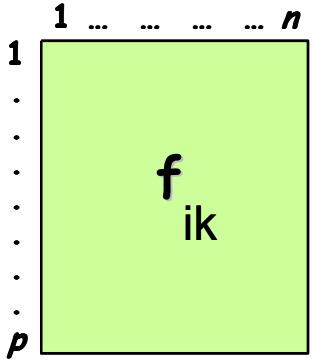
Perché raggruppare su due vie?

Un buon punto di partenza

"**In maniera abbastanza sorprendente, anche se si è interessati a raggruppare soltanto rispetto a una via della tabella di contingenza, quando si lavora su dati sparsi e altamente dimensionali, risulta proficuo il ricorso al co-clustering**"
(Dhillon, et al., 2003)

Tabella di contingenza

Nell'analisi dei dati testuali ci si riferisce di solito alla tabella lessicale come ad una tabella di contingenza (Lebart, Salem, Berry, 1998)



il cui generico elemento f_{ik} è dato dalla frequenza dell'i-esimo termine nel k - esimo documento

↓

tabella di contingenza

N. B. Alternativamente la si considera una matrice individui (documenti)-variabili (termini) one, al fine di modellizzare il comportamento lessicale

Peso computazionale

Con immense basi di dati il ricalcolo della matrice di distanze ad ogni passo (anche in algortmi k-medie) è estremamente oneroso

La scelta di un criterio da ottimizzare durante la procedura di cluster riduce il problema , essendo necessario ad ogni passo ricalcolare soltanto l'indice

Inoltre, alle volte è possibile introdurre delle considerazioni teorico sul comportamento dell'indice che possono rappresentare una ulteriore riduzione del peso computazionale

Il problema del ***clustering in una tabella di contingenza*** *è peculiare, poiché non ci troviamo nella situazione classica per cui si aggregano individui in gruppi omogenei al proprio interno ed eterogenei ve rso l'esterno, sulla base di una qualche misura derivante dai valori a ssunti dalle variabili che li descrivono*

Del resto, l'obiettivo principale che ci si pone è quello di ottenere ***gruppi di documenti****, poiché la questione di raggruppare le parole viene di regola risolta sulla base di informazione esperta (linguisti, esperti d el contesto)*

I metodi deterministici di co-clustering

In letteratura (Van Mechelen, et al., 2004), sono state individuate diverse famiglie di algoritmi di co-clustering

Qui seguiamo l'approccio deterministico, che non necessita di assunzioni distribuzionali

I metodi deterministici sono di solito block modelling

Sono basati sull'assunzione che dopo un appropriata permutazione di righe e colonne della matrice iniziale

I dati prenderanno una forma riconducibile ad una matrice a blocchi diagonali

L'obiettivo è quindi quello di ricostruire una opportuna matrice a blocchi diagonali con la minor perdita di informazione

Lavorando con tabelle di contingenza, la forza dell'associazione è classicamente misurata dal:

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(f_{ij} - f_{i.}f_{.j})^2}{f_{i.}f_{.j}} \qquad (1)$$

dove f_{ij} è la frequenza relativa della cella *ij*, $f_{i.}$ (e $f_{.j}$) sono i marginali di riga (di colonna), calcolato per l'intera tabella

Il chi-quadrato è calcolato per blocchi di celle di frequenza w_{rc} per tutti I possibili gruppi di righe e colonne:

$$\chi^2 = \sum_{r=1}^{R}\sum_{c=1}^{C}\frac{(w_{rc} - w_{r.}w_{.c})^2}{w_{r.}w_{.c}} \qquad (2)$$

La partizione ottimale sarà quella che minimizza la differe fra le equazioni (1) e (2), or, equivalentemente, massimizza il χ^2 (Greenacre, 1988) in (2)

Il τ_b di Goodman&Kruskal

Enfatizzando la causalità fra le dimensioni insito in una procedura di cluster abbiamo proposto un indice di predittività come criterio di una procedura di co-clustering

Il $\overline{\ }_b$ misura quanto la conoscenza della variabile in colonna aiuta nel predire il verificarsi delle categorie della variabile in riga

La nostra proposta (Balbi, *et al.* 2010) si basa sulla riduzione dell'errore di predizione quando la predizione si basa sulla distribuzione condizionata, piuttosto che sulla distribuzione marginale del predittore

$$\tau_b = \frac{\sum_i \sum_j f_{ij}^2 / f_{.j} - \sum_j f_{i.}^2}{1 - \sum_i f_{i.}^2} = \frac{\sum_i \sum_j \frac{(f_{ij} - f_{i.} f_{.j})^2}{f_{.j}}}{1 - \sum_i f_{i.}^2}$$

Algoritmi genetici (AG)

AG sono procedure stocastiche che richiamano meccanismi naturali quali l'ereditarietà genetica e la selezione naturale

Le potenziali soluzioni ad un problema evolvono verso una soluzione che meglio risponde al criterio interno alla funzione obiettivo

Ad ogni generazione le soluzioni migliori si riproducono, mentre quelle peggiori scompaiono

Un AG ha bisogno di definire:

- Una rappresentazione genetica dei parametri del problema
- Una funzione di valutazione (dell'adattamento)
- Degli operatori genetici (scambio, mutazione) per la popolazione

da Wikipedia (24 marzo 2014)

Algoritmo genetico

da Wikipedia, l'enciclopedia libera.

Un **algoritmo genetico** è un algoritmo **euristico** ispirato al principio della **selezione** naturale ed **evoluzione biologica** teorizzato nel 1859 da Charles Darwin

L'aggettivo "genetico" deriva dal fatto che il modello evolutivo darwiniano trova spiegazioni nella branca della biologia detta genetica e dal fatto che *gli algoritmi genetici attuano dei meccanismi concettualmente simili a quelli dei processi biochimici* scoperti da questa scienza

In sintesi si può dire che gli **algoritmi genetici** consistono in algoritmi che permettono di *valutare delle soluzioni di partenza* e che *ricombinandole* ed *introducendo elementi* di *disordine* sono in grado di crearne di *nuove* nel tentativo di *convergere a soluzioni ottime*

Queste tecniche vengono di norma utilizzate per tentare di risolvere problemi di ottimizzazione per i quali non si conoscono altri algoritmi efficienti di **complessità lineare** o **polinomiale**. Nonostante questo utilizzo, data la natura intrinseca di un algoritmo genetico, *non vi è modo di sapere a priori se sarà effettivamente in grado di trovare una soluzione accettabile* al problema considerato.

Gli algoritmi genetici rientrano nello studio dell'**intelligenza artificiale** e più in particolare nella branca della **computazione evolutiva**

Terminologia

Cromosoma: una delle soluzioni ad un problema considerato. Generalmente è codificata con un vettore di bit o di caratteri.

Popolazione: insieme di soluzioni relative al problema considerato.

Gene: parte di un cromosoma. Generalmente consiste in una o più parti del vettore di bit o caratteri che codificano il cromosoma.

Fitness: grado di valutazione associato ad una soluzione. La valutazione avviene in base ad una funzione appositamente progettata detta *funzione di fitness*.

Crossover: generazione di una nuova soluzione mescolando delle soluzioni esistenti.

Mutazione: alterazione casuale di una soluzione.

Schema di base

1. Generazione casuale della prima popolazione di soluzioni (cromosomi).
2. Applicazione della funzione di *fitness* alle soluzioni (cromosomi) appartenenti all'attuale popolazione.
3. Selezione delle soluzioni considerate migliori in base al risultato della funzione di fitness e della logica di selezione scelta.
4. Procedimento di crossover per generare delle soluzioni ibride a partire dalle soluzioni scelte al punto *3*.
5. Creazione di una nuova popolazione a partire dalle soluzioni identificate al punto *4*.
6. Riesecuzione della procedura a partire dal punto *2* ed utilizzando la nuova popolazione creata al punto *5*.

L'iterazione dei passi presentati permette l'evoluzione verso una soluzione ottimizzata del problema considerato

Rischio: "ottimi locali"

Tecniche dell'***elitarismo***" e delle ***mutazioni casuali***

La ***prima*** consiste in un ulteriore passo precedente al punto *3* che copia nelle nuove popolazioni anche gli individui migliori della popolazione precedente.

La ***seconda*** invece successiva al punto *4* introduce nelle soluzioni individuate delle occasionali mutazioni casuali in modo da permettere l'uscita da eventuali ricadute in ottimi locali.

Un AG per estrarre co-cluster

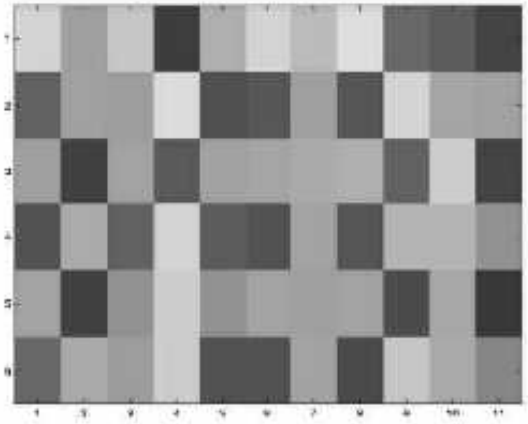
Data una tabella lessicale con *D* documenti e *K termini* si vuole ricostruire la struttura a scacchiera in cui celle con la stessa scala di grigio appartengono al co-cluster che massimizza il b

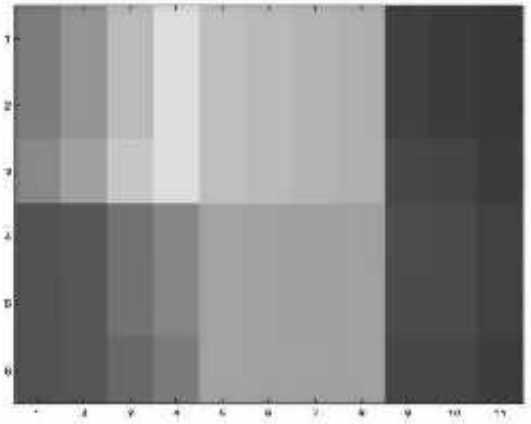
La ricerca procede per un co-cluster alla volta

Per evitare che in passi successivi si ritrovi una soluzione già trovata si crea una lista taboo per mantenere in memoria i co-cluster già trovati

⇒ **co-clusters non sovrapposti**

Graficamente :





L'algoritmo consiste dei seguenti passi:

Parametri in Input:

k (number di co-cluster da trovare)

ε (soglia di precisione)

1. **Carica la matrice dei dati e esegue, se necessario, le operazioni pre-trattamento**
2. **Inizializza casulamente la populazione (insieme candidato)**

1. **While [fitness(actual best)–fitness(best of past iteration)] > ε**
    - Calcola il criterio di 'adattamento ($_b$) su tutte le soluzioni candidate
    - Sceglie l'insieme di soluzioni che generano la nuova popolazione
    - Ottiene le nuove soluzioni applicando mutazioni e scambi delle soluzioni candidate
    - Abbandona gli elementi nella lista taboo
2. **End while**
3. **Add** la soluzione trovata **alla lista taboo**
4. **Return to step 2** fino a quando **il criterio di arresto non è soddisfatto (numero di co-clusters da trovare)**

Un esperimento su dati MEDLINE e CRANFIELD

Si tratta di due raccolte di documenti che vengono utilizzate di solito per testare nuovi algoritmi di clustering e IR

Abbiamo proceduto fondendo due sottoinsiemi di 200 documenti per ciascuna delle due collezioni

Il vocabolario congiunto consiste di 7,121 termini

Abbiamo eliminato gli apax, le parole di 1 o 2 caratteri, quelle con una frequenza superiore a 400, i numeri ed una lista comunement usata di stopwords

Alla fine abbiamo ottenuto 400 documenti and 3395 termini

Quindi la nostra matrice ha dimensioni

400 $\times$ 3395 = 1358000 celle

	Medline	**Cranfield**
C1	171	29
C2	37	163

Matrice di confusione

Riferimenti bibliografici essenziali

Agrawal R., Gehrke, J., Gunopulos D., Raghavan, P. (2005), "Automatic Subspace Clustering of High Dimensional Data", *Data Mining and Knowledge Discovery* (Springer Netherlands) 11 (1): 5–33

Balbi S. (2013) “Beyond the curse of multidimensionality: high dimensional clustering in Text Mining”, *Statistica Applicata - Italian Journal of Applied Statistics,*Vol. 22 (1)

Balbi, S., Miele, R. and Scepi, G. (2010). “Clustering of documents from a two-way viewpoint”. In: Bolasco S., Chiari I. and Giuliano L. (Eds.), *Proceedings of JADT 2010*, LED, Roma: 27-36

Berkhin, P. (2006). “A survey of clustering data mining techniques”. In: Kogan, J., Nicholas, C. and Teboulle, M. (Eds.), *Grouping Multidimensional Data Recent Advances in Clustering*, Springer, Berlin-Heidelberg: 25-71

de Falguerolle, A., Friedrich, F. and Sawitzki, G. (1997). A tribute to J.Bertin’s graphical data analysis. In: Bandilla W. and Faulbaum F. (Eds.), *SoftStat ’97, Advances in Statistical Software*, Lucius and Lucius, Stuttgart: 11-20

van Mechelen, I., Bock, H.H. and de Boeck, P. (2004). “Two-mode clustering methods: A structured overview”, *Statistical Methods in Medical Research*, (13): 363-394.