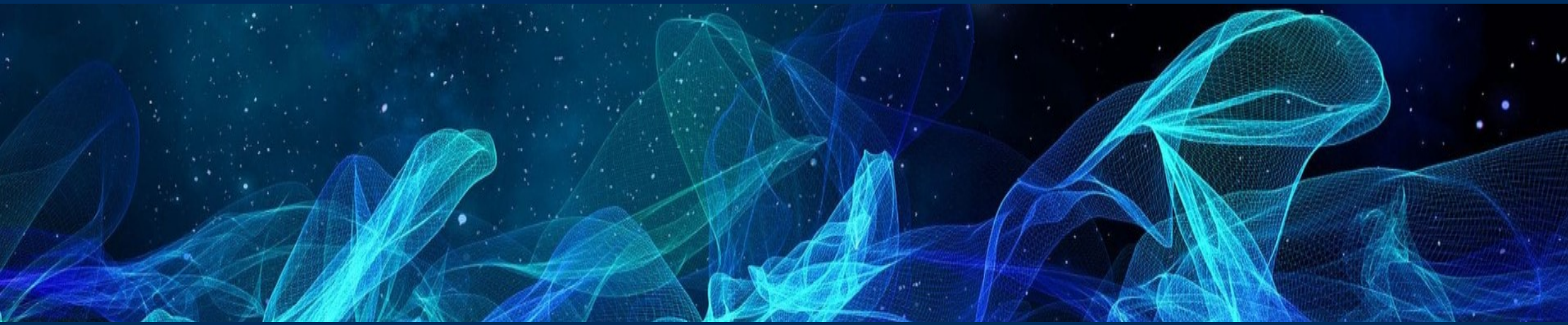
Consiglio Nazionale delle Ricerche

GenOA week 2023
Genova 24 ottobre 2023



Open public scientific data repository allow worldwide scientist to test their processing algorithms and to compare results

Diego Liberati, **Cnr-Istituto** di **Elettronica** e di **Ingegneria** dell'**Informazione** e delle T**elecomunicazioni**
diego.liberati@cnr.it

Pochi pazienti (righe) Molti attributi (colonne, espressioni geniche)

Nel lontano 1999 Golub (MIT) et al. (Science) pubblicano un metodo un po' arzigogolato per discriminare leucemie linfoidi da mieloidi sulla scorta dell'espressione genica misurata dalla nascente tecnologia dei microarrays.

I dati sono allegati all'articolo e liberamente disponibili sul sito MIT - con gli anni le buone riviste cominceranno a permettere prima, e obbligare poi, gli autori a depositarli anche come materiale supplementare all'articolo

Comincia l'era dei dati pubblici - prima tradizionalmente ciascuno era gelosissimo dei suoi dati - con i vantaggi che cercheremo di illustrare con un semplice esempio (Intelligent Data Analysis 2007)

	BioB 5_at	BioB 5_st	CreX 5_st	DapX M_at	…
Patient 1	-214	206	-118	311	…
Patient 2	-139	74	-141	134	…
Patient 3	-76	-215	84	378	…
Patient 4	-135	31	107	268	…
Patient 5	-106	252	1	118	…
Patient 6	-138	193	-1	154	…
…	…	…	…	…	…

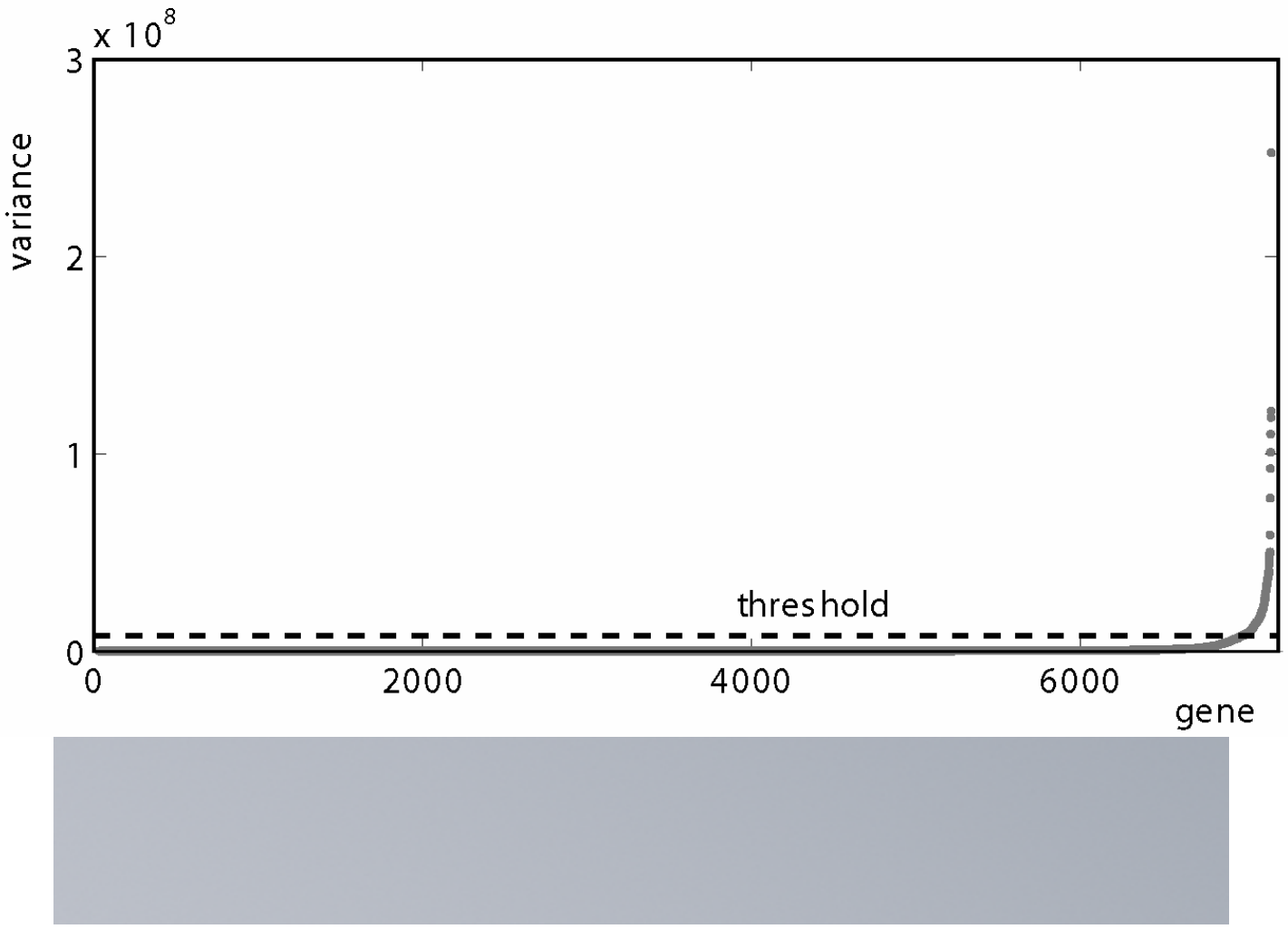
Table 1: the Leukemia data-set

Pruning: esclusione di tanti geni con piccola intervarianza di espressione

Moltissimi dei tantissimi geni pero' non mostrano sensibili variazioni tra tutti i pazienti, siano essi AML e ALL, quindi verosimilmente non saranno coinvolti nella discriminazione, li possiamo trascurare (pruning).

Mettiamo allora i geni in ordine di varianza crescente intersoggetto della loro espressione, come in figura:

Quanti (in ascissa) - o fino a che varianza (in ordinata) - trascurarne, può influire sulla qualità del risultato, non esiste un criterio univoco: il ricercatore deve avere avere intuito e/o provare e riprovare galileianamente con valori diversi.



Partizione Divisiva secondo le Direzioni Principali (PDDP)

Siccome non c'e' un unico gene tra quelli misurati capace di discriminare da solo tra AML e ALL, occorre un'analisi multivariata

In prima battuta, si ipotizza linearità (altrimenti occorrerà ricorrere per esempio a clustering binario [Muselli & Liberati, IEEE Trans CaS I, 2000)]

Si calcolano allora le componenti principali PCA della matrice multivariable dei dati dopo il pruning, e si partizionano i dati ortogonalmente alla prima componente principale, quella che esprime maggior varianza, con un inerpicano che passa per il baricentro (centroide) dei dati

PDDP clustering algorithm

Compute the centroid w of X and compute the unbiased matrix $\tilde{X} = X - ew$, $e = [1,1,\ldots,1]^T$.
Compute v, the first principal component of $\tilde{X}$.
Divide $X = [x_1, x_2, \ldots, x_N]^T$ into two subclusters X_L and X_R, according to the following rule:

$$\begin{cases} x_i \in X_L \text{ if } v^T(x_i - w) \le 0 \\ x_i \in X_R \text{ if } v^T(x_i - w) > 0 \end{cases}$$

Table 2: PDDP clustering algorithm.

Bisezione iterativa via *k*-means

Per bisecare, il popolare algoritmo k-means è semplice, ed efficace. aggrega i dati facendo si' che ciascuno di ogni gruppo sia piu' vicino a ciascun altro del gruppo stesso che ad alcuno di ogni altro gruppo.

Il numero di gruppi va definito a priori, per noi e' - iterativamente - 2, dal momento che bisechiamo

Bisecting *k-means* algorithm

Step 1. (Initialization). Select two points in the data domain space, say $c_L, c_R \in \Re^p$.

Step 2. Divide $X = [x_1, x_2, \ldots, x_N]^T$ into two sub-clusters X_L and X_R, according to the following rule:

$$\begin{cases} x_i \in X_L \text{ if } \|x_i - c_L\| \leq \|x_i - c_R\| \\ x_i \in X_R \text{ if } \|x_i - c_L\| > \|x_i - c_R\| \end{cases}$$

Step 3. Compute the centroids of X_L and X_R, w_L and w_R.

Step 4. If $w_L = c_L$ and $w_R = c_R$, stop.
Otherwise, let $c_L \leftarrow w_L, c_R \leftarrow w_R$ and go to Step 2.

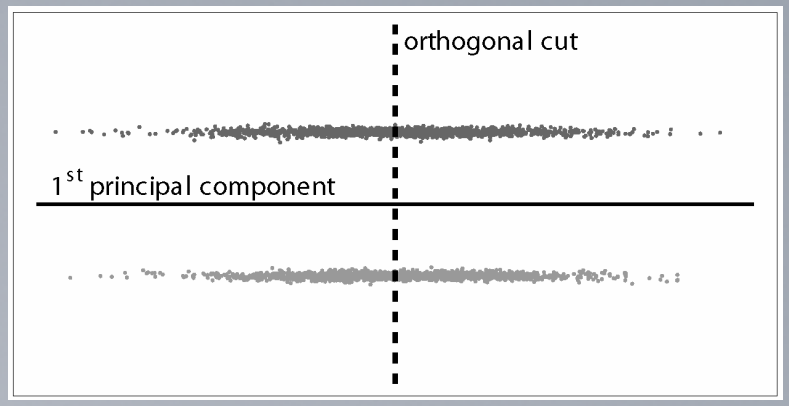
Table 3: Bisecting *k-means* algorithm.

Scelta della componente principale ortogonalmente alla quale bisecare

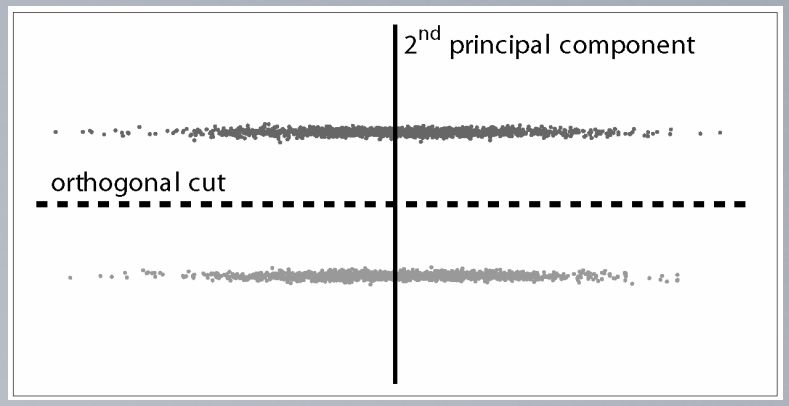
I nostri dati sono si a correlazione lineare, ma presentano una curiosa “patologia” illustrata in figura: i due gruppi di pazienti sono distribuiti in due nuvole che la prima componente principale non discrimina, la seconda in questo caso si

In generale, essendo la gerarchia della PCA a varianza decrescente da quella definita prima a quella definita ultima (dove il numero e’ il rango della matrice di correlazione) si prende la prima PC che non presenti la patologia in figura

Per noi è la seconda



(a) Cut orthogonal to the first principal component. Data aggregations are misclassified.



(b) Cut orthogonal to the second principal component. Data are correctly clustered.

Fig. 2: A data configuration where bisection should be performed orthogonally to the second principal component.

Anche l’MIT sbaglia (come per l’eccentrico nel registratore EKG analongico anni 80 (Computers Biomedical Research 1986)

Ora proiettiamo i dati nello spazio tridimensionale delle prime 3 componenti principali: il gruppo dei cerchietti piu’ in primo piano sulla destra e’ chiaramente separato da quello delle crocette in secondo piano a sinistra dalla seconda componente principale (in profondità)

Si tratta ora di verificare che in ciascun gruppo di siano tutti e soli i pazienti di una delll due classi

Uno sembrava mal classificato dal nostro algoritmo (che non sarebbe poi cosi’ male) ma si scopri’ poi che in realtà si erano sbagliati ad etichettarlo: abbiamo addirittura scoperto l’errore, discriminando con un algoritmo semplice meglio di loro con algroritmo complicato

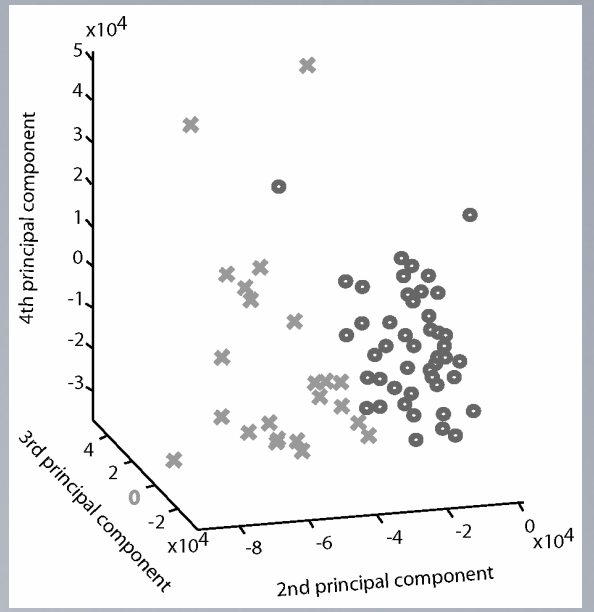


Fig. 3: PDDP + *k-means* data partition (“x”=first cluster, “o”=second cluster) .

Ogni componente principale è combinazione lineare di tutte le variabili originarie, con pesi decrescenti

Nel nostro caso, si scopre che la seconda continua.a partizionare esattamente se oltre al pruning dei geni a poca varianza operiamo un ulterior shrinking, eliminando i non necessari

Ne restano cosi' solo 8: mica male per tentare di capire cosa c'e' di diverso tra le due forme di leucemia: lasciamo tale opera ai competenti

Pero' tra gli 8 ce n'e' almeno 1 che sembra necessario alla discriminazione ma non era noto esserlo: hint, ulteriore successo!

The 8 genes able to discriminate between AML and ALL

1. FTL Ferritin, light polypeptide M11147_at
2. MPO Myeloperoxidase M19507_at
3. CST3 Cystatin C (amyloid angiopaty and cerebral hemorrage) M27892_at
4. Azurocidin gene M96326_ rna1_at
5. GPX1 Glutathione peroxidase 1 Y00433_at
6. INTERLEUKIN-8 PRECURSOR Y00787_s_at
7. VIM Vimentin Z19554_s_at
8. GB DEF Cystic fibrosis antigen mRNA M26311_s_at

Table 4: The 8 genes able to discriminate between AML and ALL

Diagnosi differenziale e piu’ profondo insight

Se guardiamo l’andamento dell’espressione genica (in ordinata) secondo i pazienti (in ascissa, con i Linfoidi tutti a sinistra di ciascun grafico e i Mieloidi a destra) troviamo conferma che nessuno - nemmeno di questo ridottissimo sottoinsieme - sarebbe in grado di discriminare da solo tra le due forme di leucemia.

L’insieme degli 8 sì, permettendoci:

- di etichettare subito un eventuale nuovo soggetto i cui fenotipi non segnalino L o M

- di capire meglio cosa non funziona in ciascuna, favorendo sia la comprensione che la diagnosi differenziale

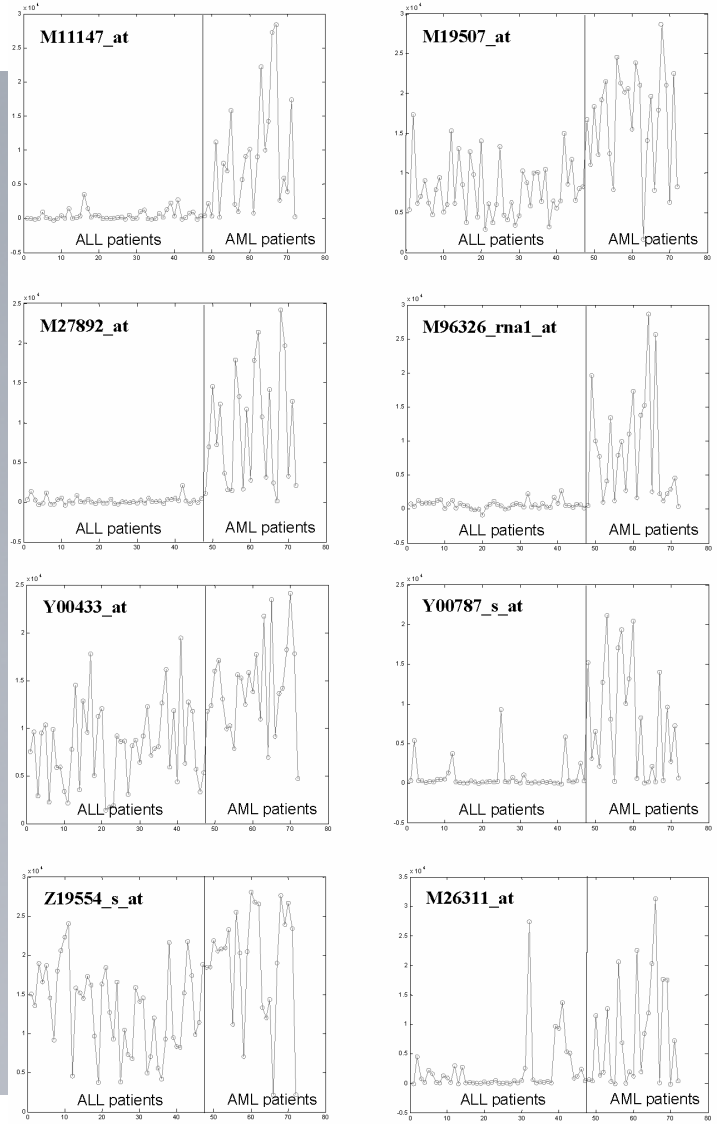


Fig. 4: expression values for the genes classifying Leukaemia patients

Bittanti, Garatti, Liberati, Maffezzoli: Intelligent Data Analysis, 2007

Grazie per l'attenzione: domande? - arrivederci!

Diego Liberati, **Cnr-Istituto** di **Elettronica** e di **Ingegneria** dell'**Informazione** e delle T**elecomunicazioni**
Diego.Liberati@cnr.it