# Introduzione

L'obiettivo di questo lavoro è quello di illustrare il clustering spettrale e di vederne alcune applicazioni nel campo della bioinformatica. In particolare nel primo capitolo viene studiato il problema di ottimizzazione discreta che porta alla derivazione degli algoritmi di clustering spettrale, mentre nel secondo applichiamo questa tecnica ad un training set ottenuto da esperimenti di microarray su oligonucleotidi di pazienti malati di leucemia, al fine di raggrupparli nelle tre categorie principali della malattia.
In generale il clustering serve a stimare le classi di appartenenza del training set con considerazioni di carattere geometrico, guardando la distribuzione e la densità dei dati in ingresso. L'obiettivo di queste tecniche è quindi quello di minimizzare la distanza degli oggetti di un gruppo e massimizzare la distanza tra gruppi diversi. Dal momento che i problemi di Clustering hanno spesso un elevato numero di dati c'é bisogno di un algoritmo efficiente. Questo algoritmo viene trovato per vie algebriche e in particolare, nel clustering spettrale, l'algoritmo si basa sul calcolo del secondo autovettore del *Laplaciano*.
Rispetto ai clustering partizionali e a quelli gerarchici, il clustering spettrale presenta diversi vantaggi. Innanzitutto la semplicità di implementazione, per la quale è sufficiente una libreria di algebra lineare, inoltre i risultati superano numerose difficoltà che si presentano con le altre tecniche.
La *teoria dei Grafi* è la teoria che sta sotto il clustering spettrale: il training set viene considerato come un grafo pesato i cui i vertici rappresentano gli oggetti del training set mentre gli archi corrispondono ai pesi di somiglianza tra i vertici che essi collegano. Il segreto di questa tecnica consiste nel partizionare questo grafo tramite il cosidetto 'taglio'. Il problema del 'taglio' viene risolto attraverso il *Laplaciano*.

# Capitolo 1

# Formulazione discreta e rilassamento

Partendo da una formulazione discreta, vediamo come impostare il problema di ottimizzazione che porta alla derivazione degli algoritmi di clustering spettrale.
Supponiamo di avere un set di oggetti, etichettati 1,2,3,...,N, indichiamo con $W \in \mathbf{R}^{N\times N}$ la matrice simmetrica dei pesi di somiglianza dove $w_{ij}$ rappresenta il peso di somiglianza dell'oggetto $i$ rispetto all'oggetto $j$ . $D \in \mathbf{R}^{N\times N}$ è la matrice diagonale con entrata diagonale pari a $d_i := \sum_{j=1}^{N} w_{ij}$. A questo punto posso definire il *Laplaciano* come la matrice $D-W$ e il *Laplaciano normalizzato* come la matrice $D^{-\frac{1}{2}}(D-W)D^{\frac{1}{2}}$. Dal teorema di Gershgorin segue che il *Laplaciano* ha autovalori non negativi $0 = \mu_1 < \mu_2 \leq \mu_3 \leq ... \leq \mu_N$ e corrispondenti autovettori reciprocamente ortonormali $\mathbf{w}^{[1]}, \mathbf{w}^{[2]}, ..., \mathbf{w}^{[N]}$, e noi siamo interessati a calcolare il relativo *vettore di Fielder* $D^{-\frac{1}{2}}\mathbf{w}^{[2]}$.
Esistono numerose altre applicazioni della *teoria dei Grafi* nel campo della bioinformatica e come vediamo in [1] e [2] sono stati implementati dei toolbox in matlab per l'imaging del connettoma e per la FC-NIRS.
In questo lavoro prenderemo in considerazione solo il *Laplaciano normalizzato* perchè l'algoritmo spettrale normalizzato è di gran lunga superiore alla versione non normalizzata nel rivelare informazioni biologicamente rilevanti. Tuttavia, come vediamo in [3], in maniera del tutto analoga si possono travare anche degli algoritmi non normalizzati a partire dal semplice *Laplaciano*.
Gli N oggetti rappresentano i vertici del nostro grafo, mentre i pesi $w_{ij}$ rappresentano

gli archi. L'obiettivo è quello di partizionare il grafo in due insiemi disgiunti A e B, usando il vettore indicatore **y**. Nel caso in cui $y_i = -\frac{1}{2}$ il vertice i è in A, mentre se $y_i = \frac{1}{2}$ allora il vertice i è in B. Una formulazione discreta per bi-partizionare il nostro grafo è la seguente:

$$\min_{i\in A, j\in B} \sum_{i,j} w_{ij} = \min_{\substack{y_i\in\{-\frac{1}{2},\frac{1}{2}\} \\ |\mathbf{y}^T D\mathbf{1}|\leq\beta}} \sum_{i,j} (y_i - y_j)^2 w_{ij}$$

Il vincolo $|\mathbf{y}^T D\mathbf{1}| \leq \beta$ proposto da Shi e Malik [4] serve a controllare la differenza tra il peso dei due clusters, per evitare di risolvere il problema di minimizzazione assegnando tutti i vertici del grafo ad un solo cluster.
Grazie alla procedura di *rilassamento* riusciamo a passare da un problema discreto difficile ad un problema continuo trattabile. Il *rilassamento* consiste nell'indebolire il vincolo da $y_i \in \{-\frac{1}{2}, \frac{1}{2}\}$ a $y \in \mathbf{R}$.
In questo modo il problema di minimizzazzione diventa il seguente:

$$\min_{\substack{\mathbf{y}\in\mathbf{R^N} \\ |\mathbf{y}^T D\mathbf{1}|\leq\frac{\beta}{\sqrt{\theta N}} \\ \mathbf{y}^T D\mathbf{y}=\mathbf{1}}} \sum_{i,j} (y_i - y_j)^2 w_{ij}$$

Nel passare da $y_i \in \{-\frac{1}{2}, \frac{1}{2}\}$ a $y_i \in \mathbf{R}$ dobbiamo occuparci del ridimensionamento del problema e lo facciamo dividendo le componenti $y_i$ per $\frac{1}{\sqrt{\theta N}}$. Inoltre il vincolo $\mathbf{y}^T D\mathbf{y} = \mathbf{1}$ evita di generare una soluzione sbilanciata in cui il singolo "outlier" *i* viene separato dal resto del gruppo.
Il problema precedente, cosi come quello non normalizzato, può essere risolto tramite il seguente teorema, che è una variazione del teorema di Rayleigh Ritz (teorema 4.2.2, [5]):

**Teorema 1.** *Sia* $A \in \mathbf{R}^{N\times N}$ *una matrice simmetrica con autovalori ordinati* $\nu_1 < \nu_2 \leq \ldots \leq \nu_N$ *e corrispondenti autovettori reciprocamente ortonormali* $\boldsymbol{x}^{[1]}, \boldsymbol{x}^{[2]}, \ldots, \boldsymbol{x}^{[N]}$. *Per*

$0 \leq \alpha < 1$ , *il problema di minimizzazione*

$$\min_{\substack{\boldsymbol{y}\in\mathbf{R}^N \\ |\boldsymbol{y}^T\boldsymbol{x}^{[1]}|\leq\alpha \\ \boldsymbol{y}^T\boldsymbol{y}=\boldsymbol{I}}} \boldsymbol{y}^T A\boldsymbol{y}$$

*ha come soluzione* $\boldsymbol{y} = \alpha\boldsymbol{x}^{[1]} + \sqrt{1-\alpha^2}\boldsymbol{x}^{[2]}$.

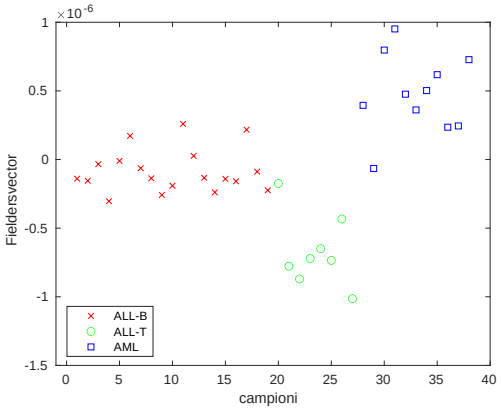
**Corollario 1.** *Per* $0 \leq \beta < \sqrt{\theta N}||D^{\frac{1}{2}}\mathbf{1}||_2$ *il problema rilassato ha la seguente soluzione:*
$\boldsymbol{y}=\frac{\beta}{\sqrt{\theta N}||D^{\frac{1}{2}}\mathbf{1}||_2^2}\mathbf{1}+\sqrt{1-\frac{\beta^2}{\theta N||D^{\frac{1}{2}}\mathbf{1}||_2^2}}D^{-\frac{1}{2}}\mathbf{w}^{[2]}$.

Il termine che moltiplica il vettore $\mathbf{1}$ non ha rilevanza nel problema del clustering, quindi il problema rilassato si riduce al calcolo del *vettore di Fielder* $D^{-\frac{1}{2}}\mathbf{w}^{[2]}$.
Possiamo anche notare che se nella formulazione discreta il problema era molto sensibile alla soglia di bilanciamento $\beta$, in questo caso, dopo il rilassamento del problema, la soluzione è completamente insensibile alla scelta della soglia.

# Capitolo 2

# Risultati degli algoritmi spettrali

In questo capitolo applichiamo gli algoritmi spettrali su dati reali, ottenuti da un esperimento di microarray su oligonucleotidi di pazienti malati di leucemia.
Il set di dati iniziale contiene campioni di midollo osseo relativi a 38 pazienti, con le relative intensità di espressione su 5000 geni. Abbiamo pre-processato questi dati sostituendo tutte le intensità di espressione minori di 20 con il valore 20. Possiamo dunque considerare i dati di interesse come una matrice $A \in \mathbf{R}^{M\times N}$,dove $a_{ij}$ rappresenta l'attivita dell' $i$-esimo gene nel $j$-esimo campione.
Nel nostro caso M=5000 geni e N=38 pazienti, di cui 27 sono classificati come ALL (leucemia linfoblastica acuta) e 11 come AML(leucemia mieloide acuta). Per quanto riguarda i campioni affetti da ALL è possibile suddividerli in base al tipo di cellule in ALL-B e ALL-T.
Costruiamo la matrice dei pesi $W = A^T A$, dove $w_{ij}$ rappresenta una misura di somiglianza tra il campione $i$ e il campione $j$. Il risultato che ci attendiamo dal clustering spettrale è quello di suddividere gli N pazienti in tre clusters che corrispondono alle tre diverse classificazioni della leucemia.
In Fig.1 possiamo vedere il comportamento del *vettore di Fielder* nel raggruppare i tre clusters. Abbiamo calcolato e graficato il comportamento di quest'ultimo con MATLAB.




In conclusione posso affermare che le tecniche di clustering spettrale possono essere utilizzate per suddividere i campioni in sottocategorie, tuttavia un ulteriore obiettivo, sempre in questa area, è quello di utilizzare queste tecniche per ottenere nuove informazioni, ad esempio per classificare il tumore di un nuovo paziente.

# Bibliografia

[1] J. Wang, X. Wang, M. Xia, X. liao, A. Evans, Y. He (2015),
*GRETNA: a graph theoretical network analysis toolbox for imaging connectomics.*

[2] J. Xu, X. Liu, J. Zhang, Z. Li, X. Wang, F. Fang, H. Niu (2015),
*FC-NIRS: A functional connectivity analysis tool for near-infrared spectroscopy data*

[3] Desmond J. Higham, Gabriela Kalna, Milla Kibble (2007),
*Spectral clustering and its use in bioinformatics.*

[4] J.Shi,J.Malik(2000),
*Normalized cuts and image segmentation.*

[5] R.A. Horn, C.R. Johnson (1985),
*Matrix Analysis*