Caratterizzazione e Clustering di Matrici di Correlazione ADNI-2 Tramite Distribuzione di Wishart

Carlo Mengucci

7 dicembre 2017

Indice

1 Introduzione

In questo estratto vengono presentati i risultati di *clustering* e *classificazione* per soggetti ADNI-2, di cui sono state analizzate le matrici di correlazione ottenute tramite elaborazione in Macrovoxel, utilizzando la *Distribuzione di Wishart* come ipotesi nulla.

Vengono inoltre presentati sinteticamente l'algoritmo e la *pipeline* di *cross-validazione* utilizzati.

2 Struttura e Utilizzo del Database

Dei 403 soggetti totali sono stati utilizzati 232 soggetti a seguito di operazioni di filtraggio e dell'individuazione di discrepanze nelle procedure di normalizzazione. E' infatti possibile riscontrare la presenza di due gruppi distinti all'interno del database, dei quali è stato preso in considerazione il più numeroso.

I 232 soggetti utilizzati per l'elaborazione finale sono stati divisi per il *training* utilizzando la sola discriminante della conversione. Sono pertanto utilizzate le labels *NC (Non-Converters)* e *AD (Alzheimer Diagnosis)* per il clustering; la presenza di soggetti per il gruppo *AD* è del 63% mentre il restante 37% dei 232 soggetti considerati appartiene al gruppo *NC*.

Ad ogni soggetto è associata una matrice di correlazione $N \times N$, $N = 549$, Ognuno degli N elementi rappresenta un *Macrovoxel* di cui è estratta la correlazione *topologica* rispetto a tutte le altre componenti del sistema. Ogni *Macrovoxel* è definito su un insieme di $3 \cdot 10^3$ Voxel.

3 Distribuzione di Wishart e Definizione delle Condizioni di Applicazione

3.1 Definizione della Distribuzione di Wishart

La distribuzione di Wishart consiste in una famiglia di distribuzioni per *matrici simmetriche definite positive*.

Def. Siano $X_1 ... X_n$ indipendenti $N_p(0, \Sigma)$ distribuiti, tali da formare una matrice di dati $p \times n$, $X = [X_1 ... X_n]$. La distribuzione di *matrici random* $p \times p$, $M = XX' = \Sigma_{i=1}^{n} X_i X_i'$ è una distribuzione di Wishart. [1]

La matrice random $M_{p \times p} = \Sigma_{i=1}^{n} X_i X_i'$ segue una distribuzione di Wishart a n gradi di libertà e *matrice di covarianza* Σ ed è definita $M \sim W_p(n, \Sigma)$. Per $n \geq p$ la *pdf* di M assume la forma [1]:

$$f(M) = \frac{1}{2^{\frac{np}{2}} \Gamma_p(\frac{n}{2}) \|\Sigma\|^{\frac{n}{2}}} \|M\|^{\frac{n-p-1}{2}} exp[-\frac{1}{2} trace(\Sigma^{-1} M)] \qquad (1)$$

La Wishart può essere interpretata come l'estensione multivariata di una distribuzione χ^2.

3.2 Definizione delle Condizioni di Applicazione

Le matrici di correlazione sono per definizione simmetriche definite positive.

[1] *Nota:* $\|\Sigma, N\| = det(\Sigma, M)$

Il numero n di gradi di libertà del sistema è dato dal campionamento ($n = 3 \cdot 10^3$) del singolo Macrovoxel.

Utilizzando come matrice di scala la matrice data dalla media delle matrici di correlazione delle due categorie di soggetti, è possibile stimare la Wishart attesa per le categorie stesse.

Un approccio di questo tipo permetterebbe la classificazione ogni soggetto in base alla propria distanza dalle distribuzioni rappresentative delle categorie, in termini di *LogPDF*, come definito dalla eq.(2):

$$score_{subj} = logP_W(\Sigma_{subj} \mid n, \hat{\Sigma}_{AD}) - logP_W(\Sigma_{subj} \mid n, \hat{\Sigma}_{NC}) \tag{2}$$

dove Σ_{subj} rappresenta la matrice del singolo soggetto e $\hat{\Sigma}_{AD,NC}$ è la matrice media di categoria.

L'algoritmo implementato rende possibile la stima dello score, relativo alla distanza dalle distribuzioni attese per le categorie, *feature* per *feature*. Per ogni paziente viene infatti calcolato lo *score* definito dall' eq. (2) eliminando di iterazione in iterazione una diversa componente del sistema; in questo modo è possibile stimare il peso che ogni *feature* possiede all'interno del sistema stesso.

Nel caso presente le *features* sono le componenti della matrice di correlazione associata al soggetto, ossia i 549 *Macrovoxel* rappresentativi del sistema.

4 Pipeline

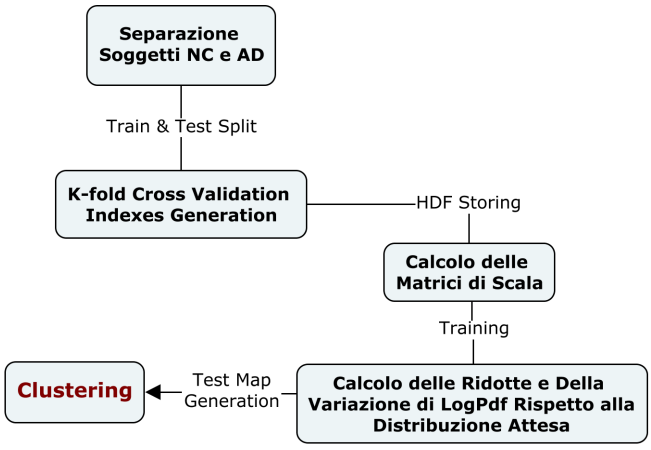

Figura 1: *Schema di Workflow della pipeline di elaborazione*

In figura (1) è riportato lo schema di funzionamento dell'intera pipeline. Viene testata la capacità di discriminare tra NC e AD attraverso un processo di *supervised learning*.

A seguito di una divisione iniziale nelle due classi, vengono separati i soggetti in *train* e *test* secondo il classico split $90\% - 10\%$ per ogni batch della *K-fold cross-validation*.

Da ogni batch di training è calcolata la matrice di scala da cui è generata la distribuzione che viene confrontata con il corrispettivo batch di test.

Per ogni soggetto dei batch di test è infine generato un vettore contenente gli *scores* relativi alle singole features, i quali vanno a comporre le mappe finali su cui vengono effettuati clustering e classificazione.

Per il clustering è stata utilizzata una metrica di tipo *City Block* mentre la separazione è stata effettuata utilizzando una SVM a *Kernel Lineare* e parametro $C = 1$. Gli score di classificazione sono quindi stimati tramite una cross-validazione 10-fold stratificata.

5 Risultati

Come riportato in figura (2), l'utilizzo degli scores relativi alle singole features permette non solo una buona separazione fra i gruppi NC e AD, ma fornisce anche informazioni su quali siano i gruppi di features più determinanti nella classificazione dei soggetti.

In figura (3) sono invece riportate le distribuzioni relative alla somma degli scores dei singoli pazienti. L'informazione viene cioè ridotta ad un singolo score per paziente; ciò non inficia tuttavia la capacità di discriminazione rappresentata dalla forte distinzione tra i due picchi.

Questo risultato è confermato anche dalla performance della SVM che fornisce un'*accuracy* di classificazione del 100%.

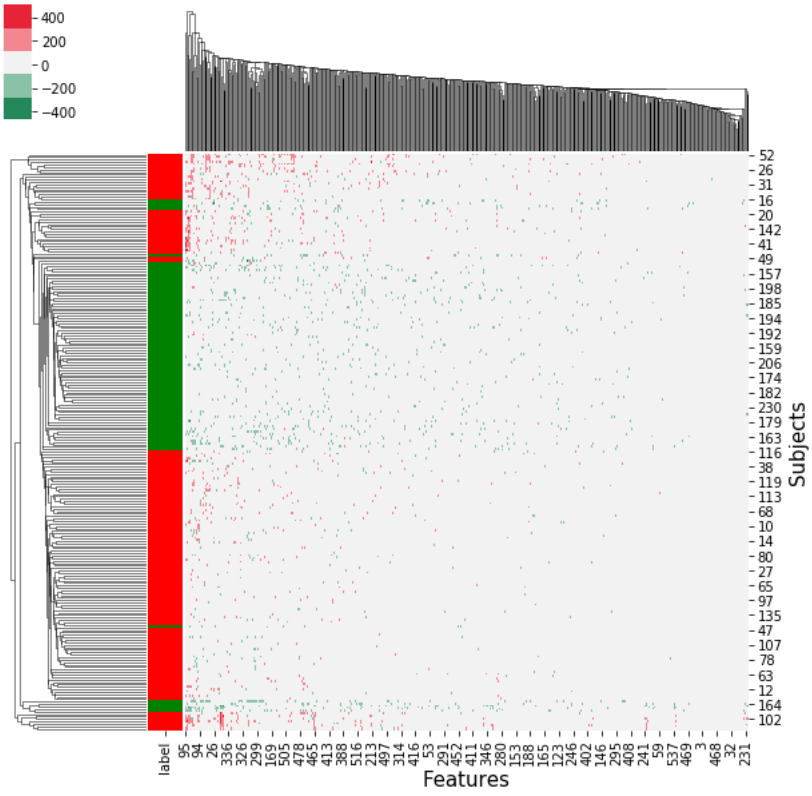

Figura 2: *Risultato del Clustering. Il gradiente di colore è relativo alla significatività delle Features nel stabilire l'appartenenza all'una o all'altra categoria. In verde sono rappresentati i soggetti NC ed in rosso i soggetti AD.*

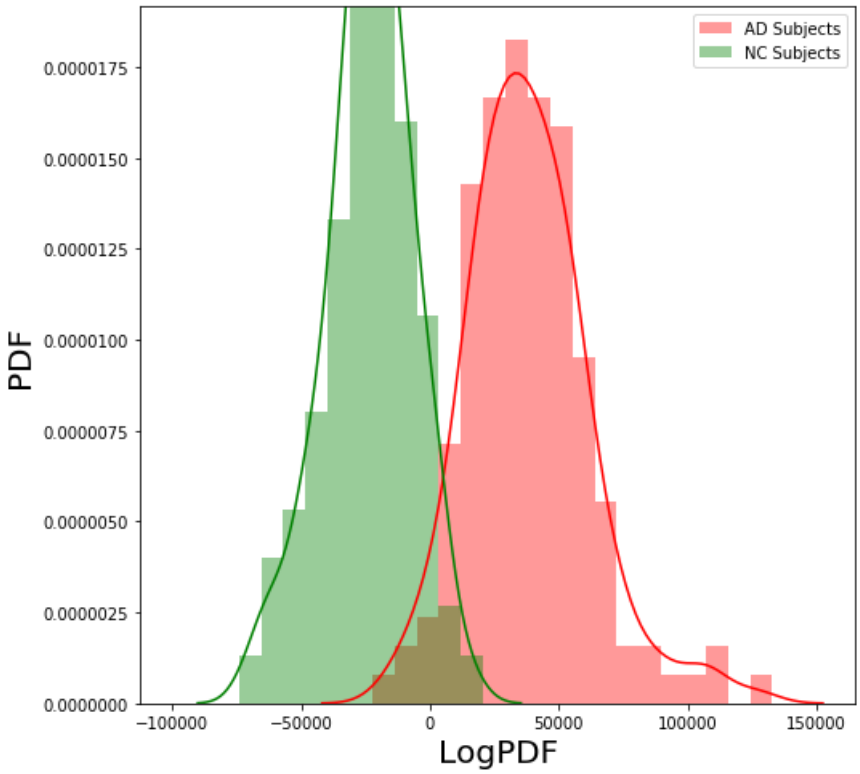

Figura 3: *Distribuzioni della somma degli scores per paziente. Le classi sono ben distinte anche riducendo ad un unico score l'informazione sui soggetti.*

Riferimenti bibliografici

[1] Hardle, Wolfgang and Leopold Simar
Applied Multivariate Statistical Analysis
Heidelberg: Springer Berlin Heidelberg, 2012