Big Data in Health Care

Federico Moiraghi - 799735 & Pranav Kasela - 846965

A.A. 2019/2020

Sommario

Obiettivo del presente Progetto è di fornire due modelli predittivi che riescano a riconoscere i tumori omogenei da quelli eterogenei. Infatti l'eterogeneità del tumore rappresenta una difficoltà aggiuntiva nella fase di trattamento, rendendolo maggiormente resistente alle cure. Il primo modello, *supervised*, sarà facilmente interpretabile da un esperto di dominio, in modo tale da supportare le sue decisioni senza sostituirsi completamente ad esso. Tale modello sarà poi confrontato con uno *unsupervised*, che sottolinea le analogie tra i singoli casi.

Indice

1 Introduzione

Si è deciso di sviluppare il presente Progetto coi linguaggi di programmazione Python ed R, data la forte crescita del loro uso sia in ambiente accademico che produttivo. Il primo è usato soprattutto per l'estrazione delle *features* dalle immagini, grazie alla libreria `pyradiomics` che offre numerosi algoritmi, mentre il secondo per l'analisi dei dati, data l'ampia scelta di modelli di *machine learning*.

Caricate le immagini (si ha un esempio con figura 1), si nota che queste rappresentano la lesione già segmentata. Non si ritiene dunque necessaria alcuna forma particolare di *pre-processing* sull'immagine.

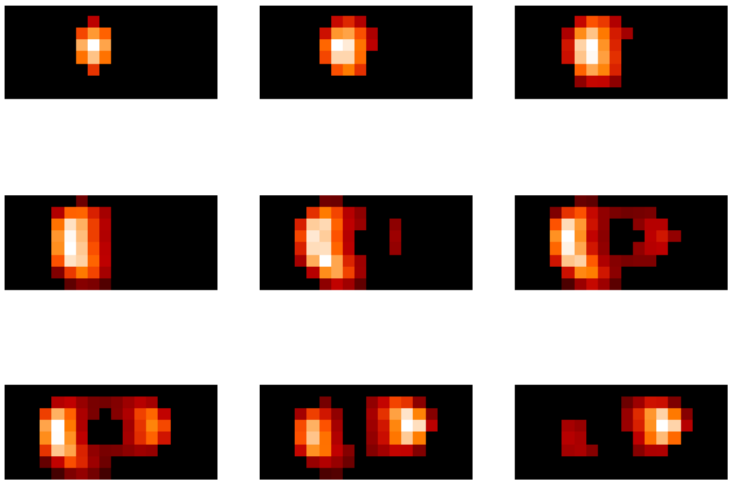

Figura 1: Esempio di immagine. Essendo una figura tridimensionale, si rappresenta la profondità con più immagini.

Mentre le dimensioni dei *voxel* sono fisse per tutti i file, l'immagine è ritagliata sulla lesione in modo specifico, tralasciando le parti adiacenti: le dimensioni variano in base alla dimensione della lesione stessa.

	voxel size
x	2.734
y	2.734
z	2.734

Dunque sarà importante estrarre delle *features* che non dipendano dalla dimensione dell'immagine ma tengano conto di possibili variazioni. Questo approccio comporta una serie di vantaggi, primo tra tutti la modularità del *workflow*: è possibile così prevedere la variabile risposta avendo a disposizione sia un'immagine già segmentata sia effettuando la segmentazione *on-the-fly* tramite semplici algoritmi a soglia, riducendo potenzialmente il tempo della diagnosi.

2 Estrazione delle *features*

L'estrazione delle *features* mappa le immagini in uno spazio di dimensionalità molto minore e di dimensioni fisse, rendendo più semplice l'analisi dato il numero esiguo di dati a disposizione. Infatti, un qualsiasi algoritmo di *machine learning* ha bisogno di un numero significativo di dati per "apprendere" in modo *data-driven* cosa utilizzare nell'analisi.

Per selezionare le *features* da utilizzare, si è preso spunto da [1][1] e [2]: si usano *features* estratte direttamente dall'immagine, quali superficie della lesione o la sua sfericità, accompagnate da indici statistici più semplici quali i momenti di ordine dal primo fino al quarto (media, varianza, asimmetria e curtosi) dei valori dei singoli *voxel* (tabella 1).

VoxelNum	Maximum3DDiameter	MajorAxisLength	Sphericity
-1.337	-4.107	-3.071	1.116
-0.036	0.727	0.35	-0.381
0.458	0.118	-0.106	0.956
0.192	0.745	0.71	-0.546
-0.654	0.14	0.664	-0.124
-0.329	-0.508	-0.692	0.607

Tabella 1: Esempio di *features* estratte per le singole immagini.

Tuttavia il numero di *features* estratte è ancora elevato rispetto al numero di dati a disposizione (la matrice di *input* ha dimensioni 44×24). Si effettua dunque una prima selezione delle *features* tramite il test di Mann-Whitney, equivalente non parametrico del t-test (i risultati sono riassunti nella tabella 2), per escludere le variabili la cui significatività, prese singolarmente, è minore del 5%.

[1]Gli autori usano i primi quattro momenti per stimare la differenza di eterogeneità di tumori alla cervicale nel tempo, a seguito di un trattamento.

Maximum3DDiameter	0.011
MajorAxisLength	0.012
Sphericity	0.003
MinorAxisLength	0.003
SurfaceArea	0.016
Kurtosis	0.002
Maximum	0.008
Skewness	0.04
Variance	0.025
Contrast	0.02
$\text{ngtdm}_{\text{Coarseness}}$	0.034

Tabella 2: p-value delle varibili accetate dal test di Mann-Whitney.

Effettuata questa prima cernita, si riduce ulteriormente il numero di *features*, in modo tale da evitare multi-collinearità tra le variabili, rispettando così le premesse del modello lineare.

Dal correlogramma (figura 2) si deduce quali variabili escludere (`Maximum`, `Variance`, `Maximum3DDiameter`, `MinorAxisLength`, `Contrast` e `Sphericity`): la matrice risultante ha una dimensionalità ridotta (44×5), adeguata per la costruzione del modello.

3 Modello *supervised*

Essendo la variabile risposta binaria (tumore *omogeneo* o *eterogeneo*, rispettivamente 0 o 1), e volendo costruire un modello facilmente interpretabile per un esperto di dominio, si effettua una semplice regressione logistica.

La selezione delle *features* è effettuata tramite procedimento *stepwise* usando l'indice BIC[2], con possibilità di re-immissione. Il numero di variabili significative si riduce quindi a tre: `SurfaceArea`, `Kurtosis` e `Skewness` (riassunti nella tabella 3 coi rispettivi p-value).

Le prestazioni del modello sono calcolate col sistema *cross validation*, effettuando 30 iterazioni casuali dividendo i dati 80% *train set* e 20% *test set*, così da avere stime robuste dei parametri e un intervallo di confidenza sufficientemente ristretto. La media degli indici di bontà è riportata nella tabella 4 assieme al rispettivo intervallo di confidenza al 99%.

[2]L'indice BIC rispetto all'indice AIC penalizza maggiormente l'inserimento di una nuova variabile con un numero ridotto di osservazioni.

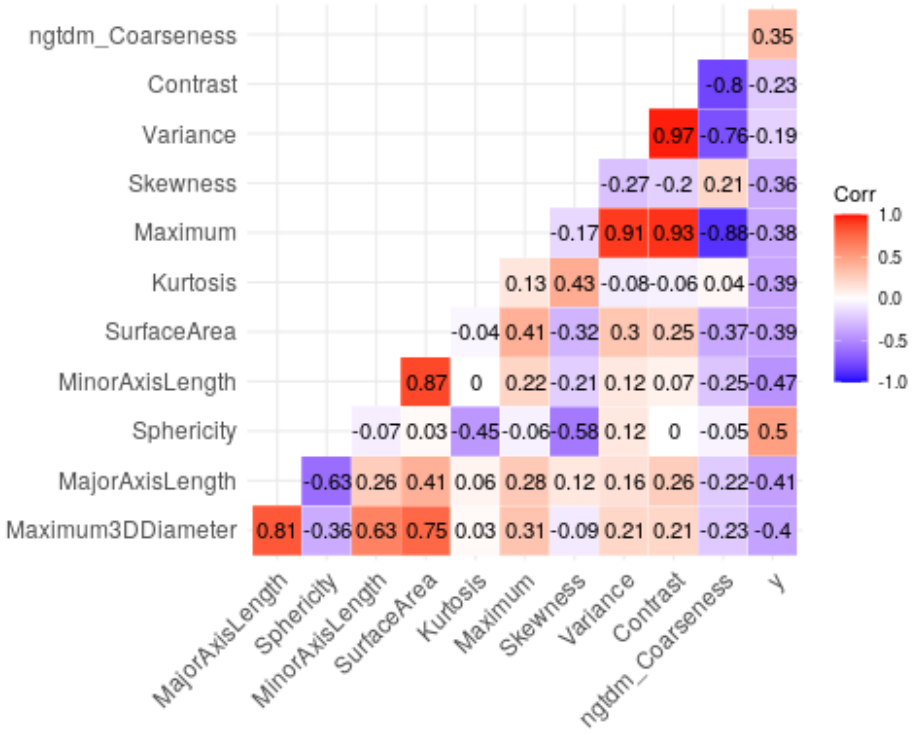

Figura 2: Correlogramma delle *features* estratte.

	Stima	p-value
(Intercept)	-4.295873	0.014587
SurfaceArea	-11.899879	0.005449
Kurtosis	-9.842963	0.008876
Skewness	-8.655905	0.007367

Tabella 3: Stima dei coefficienti del modello e loro significatività.

	average	IDC_{99}
accuracy	0.893	0.04
precision	0.932	0.054
recall	0.89	0.059
f_1	0.899	0.041

Tabella 4: Performance del modello supervisionato con intervallo di confidenza al 99%.

Si calcola quindi anche la matrice di confusione (tabella 5), e si nota che l'unico errore è stato commesso catalogando come eterogenea una variabile omogenea.

	heterogeneous	homogeneous
prediction: heterogeneous	4	1
prediction: homogeneous	0	4

Tabella 5: Matrice di confusione del modello di regressione logistica per il *test set*; sulle righe le previsioni e sulle colonne i valori reali.

4 Modello *unsupervised*

Considerando tutti i dati (quindi più informazione possibile), standardizzati, si effettua una divisione in *clusters* con l'ipotesi che sia possibile raggruppare le due tipologie di tumore.

Le immagini di tumori quindi sono collocate in uno spazio vettoriale in base al risultato della *Principal Component Analysis* (PCA): si selezionano così le prime 6 componenti, che assieme spiegano circa il 95% della varianza totale della distribuzione. Così, oltre a operare su una matrice di dimensioni ridotte, si riduce anche la quantità di rumore data dall'elevato numero di variabili (a cui si esclude la variabile risposta `y`, usata poi per calcolare la bontà del modello) spesso inutili. Dalla figura 3 infatti si vede che all'aumentare del numero di componenti considerate, la percentuale di varianza colta aumenta con un tasso decrescente: la soglia del 95% è un compromesso tra il segnale colto dal modello e la sua complessità (per i dettagli vedere la tabella 6).

Nello spazio della PCA si effettua un raggruppamento usando l'algoritmo DBScan, basato sulla densità delle osservazioni. La figura 4 suggerisce un parametro $\varepsilon = 3.5$ (con 5-NN): questa configurazione sarà usata per la costruzione del modello.

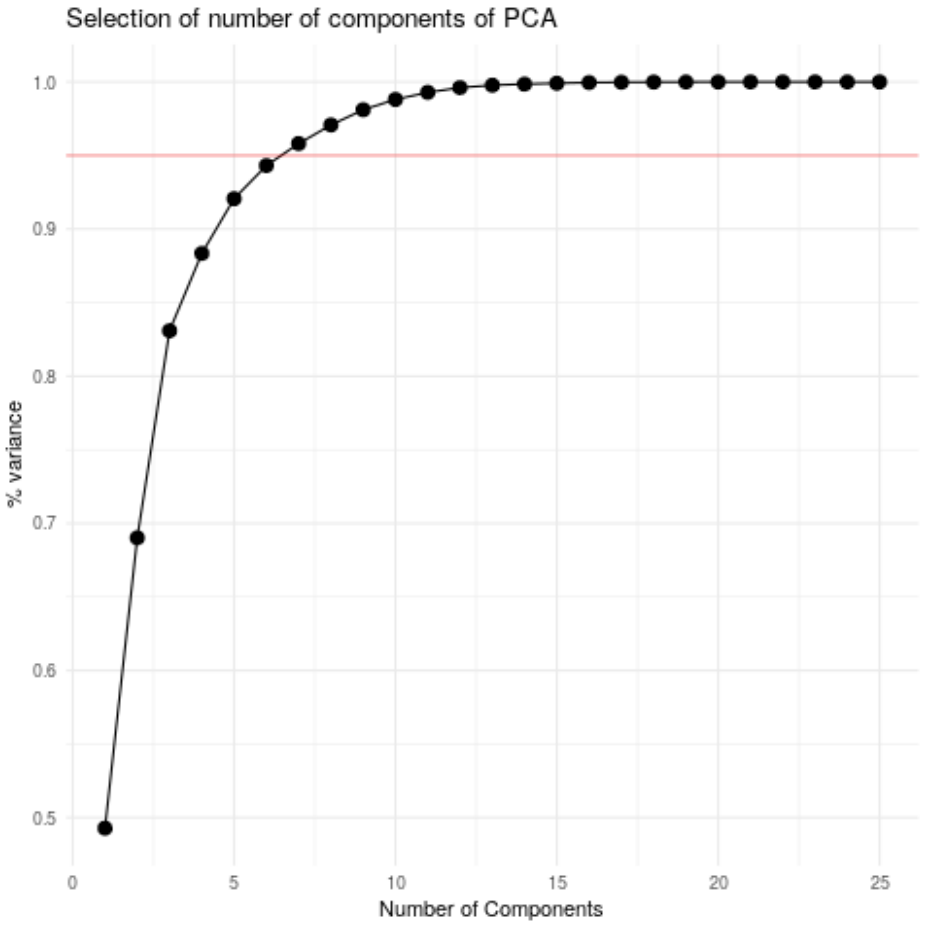

Figura 3: Andamento della varianza spiegata dal modello all'aumentare del numero di componenti della PCA.

	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	3.454	2.188	1.824	1.122	0.941	0.728
Proportion of Variance	0.497	0.199	0.139	0.052	0.037	0.022
Cumulative Proportion	0.497	0.696	0.835	0.887	0.924	0.946

Tabella 6: Alcune statistiche sulle prime componenti principali.

accuracy	0.773
precision	0.962
recall	0.735
f_1	0.833

Tabella 7: Indici di bontà per la clusterizzazione con DBScan.

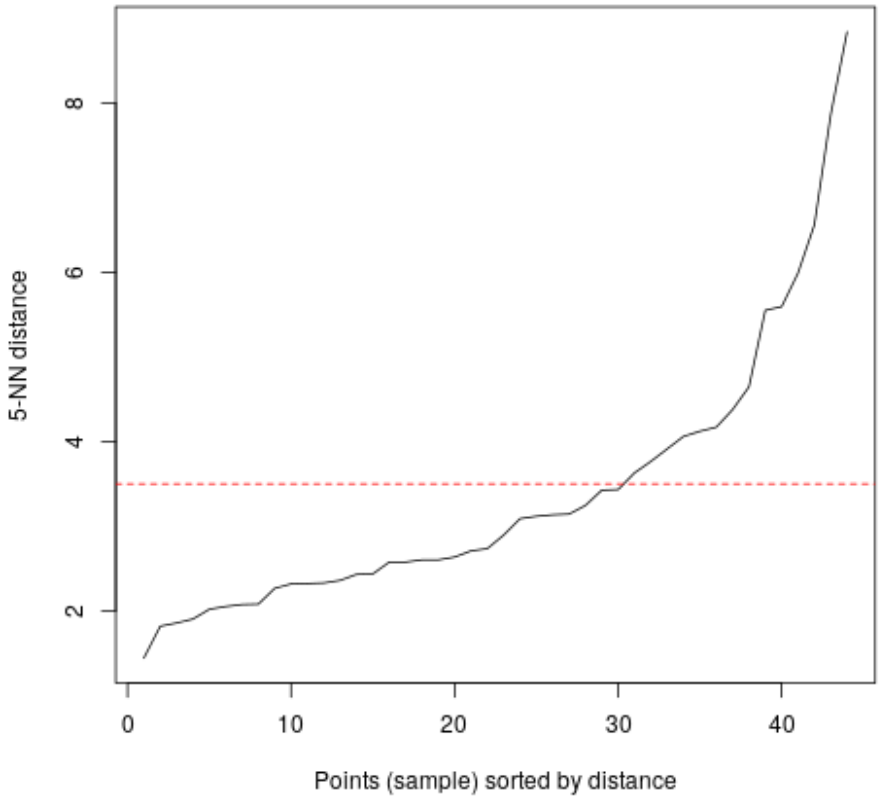

Figura 4: Scelta del valore ε per DBScan.

	heterogeneous	homogeneous
C_0	1	9
C_1	25	9

Tabella 8: Distribuzione delle immagini all'interno dei *clusters*.

Nonostante le buone *performance* del modello (riassunte nella tabella 7), si nota che il secondo *cluster* C_1 contiene un numero non indifferente di immagini omogenee. Infatti l'algoritmo è riuscito a individuare un solo *cluster* (C_0) ben definito, considerando la variabile risposta.

Si tenta quindi un altro approccio, con l'algoritmo *HK-means*, versione gerarchica del ben più noto *K-means*. L'algoritmo è quindi testato con un numero di *cluster* k da 2 a 15, calcolando per ciascuno la distanza nei gruppi (*distance between*). La figura 5 mostra graficamente il procedimento: si sceglie $k = 4$ per evitare *overfitting* dei dati, e siccome il tasso di aumento per $k > 4$ decresce fortemente. La bontà del raggruppamento invece (intesa come capacità predittiva) è riassunta nella tabella 9.

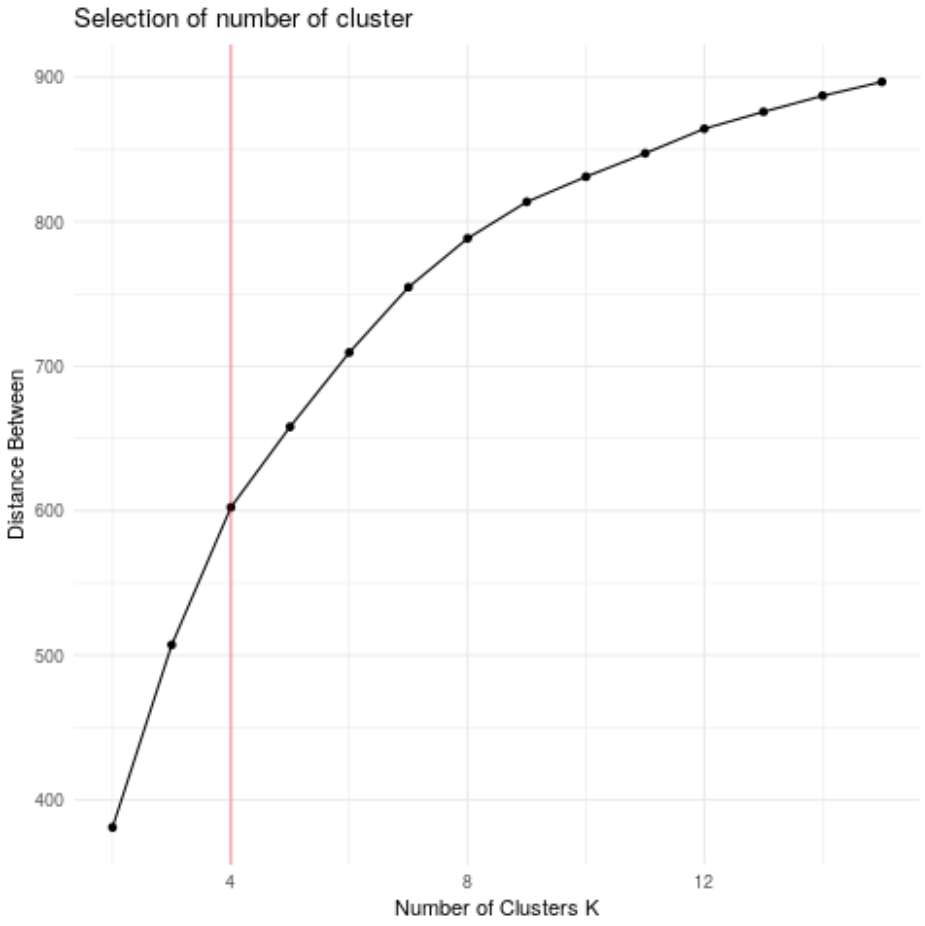

Figura 5: Variazione della distanza *between* all'aumentare del parametro k.

Nonostante le buone prestazioni, a confronto col modello DBScan non si hanno forti miglioramenti: infatti il modello precedente ha una capacità

accuracy	0.727
precision	1
recall	0.684
f_1	0.813

Tabella 9: Indici di bontà per HK-Means con $k = 4$.

	heterogeneous	homogeneous
C_1	18	8
C_2	8	4
C_3	0	4
C_4	0	2

Tabella 10: Il modello ha identificato due *cluster* definiti (C_3 e C_4), considerando la variabile risposta; mentre la maggior parte delle osservazioni è compresa nei primi due *cluster* (C_1 e C_2).

predittiva maggiore e una complessità minore. Nonostante questo, la capacità predittiva dei modelli *unsupervised*, paragonati a quello *supervised* presentato precedentemente, è nettamente inferiore. Inoltre, lavorando nello spazio delle componenti principali, l'interpretabilità del modello risulta difficile anche per un esperto di dominio.

5 Conclusioni

Con questo Progetto si è costruito un modello statistico *supervised* efficace e facilmente interpretabile da un esperto di dominio per prevedere l'eterogeneità del tumore. Per costruirlo è stato sufficiente estrapolare dalle immagini segmentate delle semplici *features*, veloci da calcolare e facili da interpretare. Si è quindi confrontato questo modello con uno *unsupervised*, confermando la superiorità del primo sia per bontà di previsione sia per facilità di interpretazione.

Per migliorare il modello si potrebbe, a livello teorico, usare un numero maggiore di dati per la stima dei parametri e per selezionare le *features* da includere; tuttavia questo non è sempre possibile in ambito medico, data la forte difficoltà e l'alto costo nell'ottenere una più grande quantità di dati.

Riferimenti bibliografici

[1] Stephen Bowen, William Yuh, Daniel Hippe, Wei Wu, Savannah Partridge, Saba Elias, Guang Jia, Zhibin Huang, George Sandison, Dennis Nelson, Michael Knopp, Simon Lo, Paul Kinahan, and Nina Mayr. Tumor radiomic heterogeneity: Multiparametric functional imaging to characterize variability and predict response following cervical cancer radiation therapy. *Journal of Magnetic Resonance Imaging*, 47, 10 2017.

[2] Francesca Gallivanone, Matteo Interlenghi, Daniela D'Ambrosio, Giuseppe Trifirò, and Isabella Castiglioni. Parameters influencing pet imaging features: a phantom study with irregular and heterogeneous synthetic lesions. *Contrast Media & Molecular Imaging*, 2018:1–12, 2018.