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*, anch'esso di facile interpretazione, 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`, mentre il secondo per l'analisi dei dati, data l'ampia disponibilità di modelli di *machine learning* già implementati.

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.

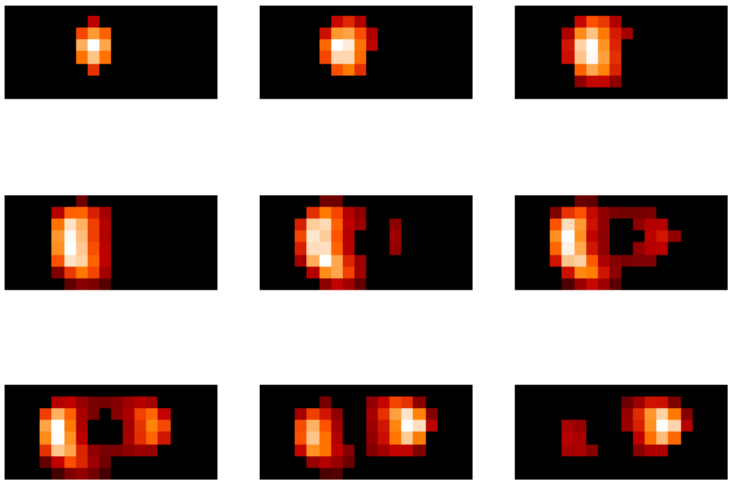

Figura 1: Esempio di immagine. Essendo una figura tridimensionale, si rappresenta la profondità con più immagini.

Mentre le dimensioni dei *voxel* sono fisse per tutti i files, 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 dal primo fino al quarto (media, varianza, asimmetria e curtosi) dei valori dei singoli *voxel* (tabella 1).

VoxelNum	Maximum3DDiameter	MajorAxisLength	Sphericity
0.485	0.356	0.028	0.787
-0.295	-0.719	-0.747	0.85
0.458	0.118	-0.106	0.956
-0.341	-0.6	-0.021	-1.781
-0.326	-0.719	-0.751	0.736
-0.739	-0.14	0.062	-1.256

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 da sole è minore del 5%.

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

[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
$\text{ngtdm}_{\text{Coarseness}}$	0.034

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

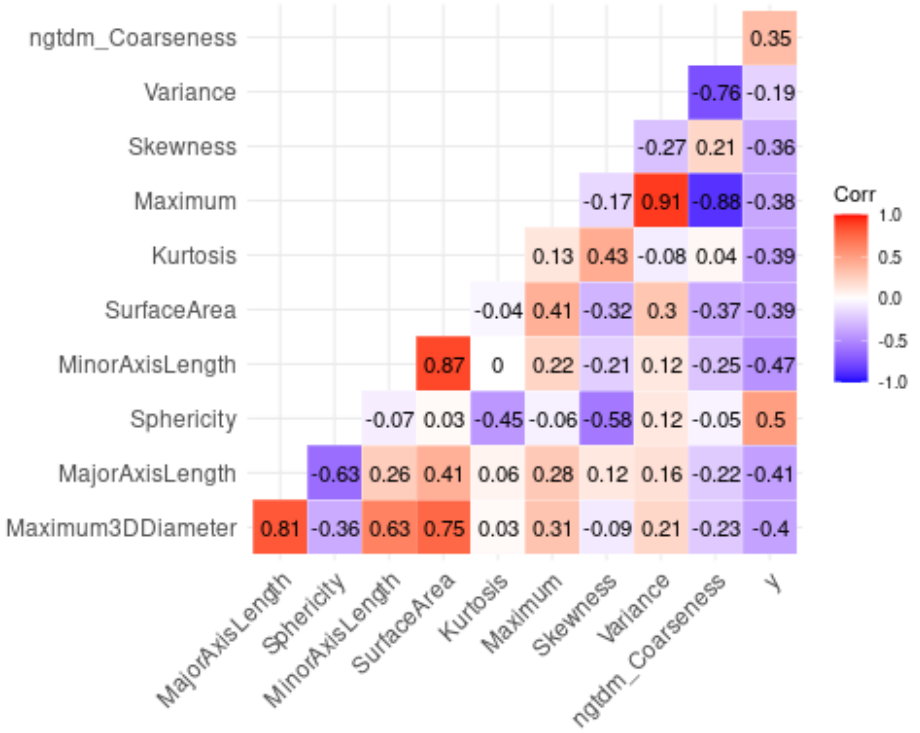

Figura 2: Correlogramma delle *features* estratte.

Dalla figura 2, si deduce quali variabili escludere (`Maximum`, `Variance`, `Maximum3DDiameter`, `MinorAxisLength`, `Contrast` e `Sphericity`): la matrice risultante ha così dimensioni 44×5, dimensionalità 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.

	average	IDC_{99}
accuracy	0.889	0.057
precision	0.932	0.054
recall	0.891	0.059
f_1	0.905	0.049

La selezione delle *features* è effettuata tramite procedimento *stepwise* (partendo dal modello pieno ed eliminando le variabili superflue, ma con la possibilità, a ogni iterazione, di reinserirle). Si è deciso di rimuovere alcune variabili a prescindere:

- `sd` (la varianza della distribuzione della lesione), in quanto fortemente correlata con `mean` (0.93) e di più difficile interpretazione;
- `area` (la superficie della lesione), in quanto la sua stima è approssimativa e risulta essere eccessivamente correlata ad altri regressori, inquinando eccessivamente la qualità dei dati;
- `sphere` (il volume della sfera equivalente), data la sua forte correlazione con la variabile `volume` (0.86) e la più difficile interpretabilità.

Avendo a disposizione pochi dati, il procedimento è effettuato con l'indice AIC, che considera la capacità di generalizzazione del modello complessivo risultante (salvo poi verificare le *performance* su un *test set* composto da dati nuovi, rappresentante circa il 20% di quelli totali). Alla fine del procedimento, il modello risultante comprende solo tre regressori (più l'intercetta), come mostrato in tabella 3.

Si noti come i coefficienti maggiormente significativi siano l'asimmetria `sk` e il volume `volume`: la probabilità che il tumore sia eterogeneo è tanto maggiore quanto più grande è la lesione e quanto più pesante è la coda positiva della distribuzione. Si può ipotizzare infatti che questa coda positiva sia

	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.

	heterogeneous	homogeneous
heterogeneous	4	0
homogeneous	0	5

Tabella 4: Matrice di confusione del modello di regressione logistica per il *test set*; sulle righe le previsioni e sulle colonne i valori reali.

costituita da sotto-componenti particolarmente aggressivi del tumore, quindi "ghiotti" di traccianti e di conseguenza maggiormente visibili nell'immagine.

Come si può notare nella matrice di confusione (figura 4), il modello ha commesso un solo errore catalogando come eterogenea una lesione omogenea.

accuracy	1
precision	1
recall	1
f_1	1

4 Modello *semi-supervised* o *unsupervised*

Riprendiamo il dataset di partenza senza alcun feature selection, e effettiamo solo una standardizzazione. Per effettuare il clustering si devono diminuire le dimensioni, non volendo usare le y (se non per il calcolo dell'accuratezza) non possiamo usare i soliti modelli di feture selection, un modo potrebbe essere effettuare una feature selection basata sulla varianza, e per fare cio' si effetua la PCA, accettando solo le PCA che spiegano circa il 95% della varianza delle variabili di partenza. Nella figura 3 si vede che 6 componenti del PCA spiegano il 95% della varianza, in questo modo si elimina anche molto rumore che e' causato dalla presenza di tante variabili (spesso inutili).

Si effettua un primo tipo di clustering basato sulla densita' (nello spazio dei PCA) usando uno dei modelli piu' usati per il clustering, il DBScan. Nella Figura 4 notiamo che eps piu' vantaggioso usando la distanza di 5-NN risulta vicino al 3.5, quindi viene scelto esso come il valore ideale, per il clustering.

Pur avendo un'accuratezza non bassa, si vede che il secondo cluster contiene un numero di immagini omogenee non trascurabili, inoltre il DBScan

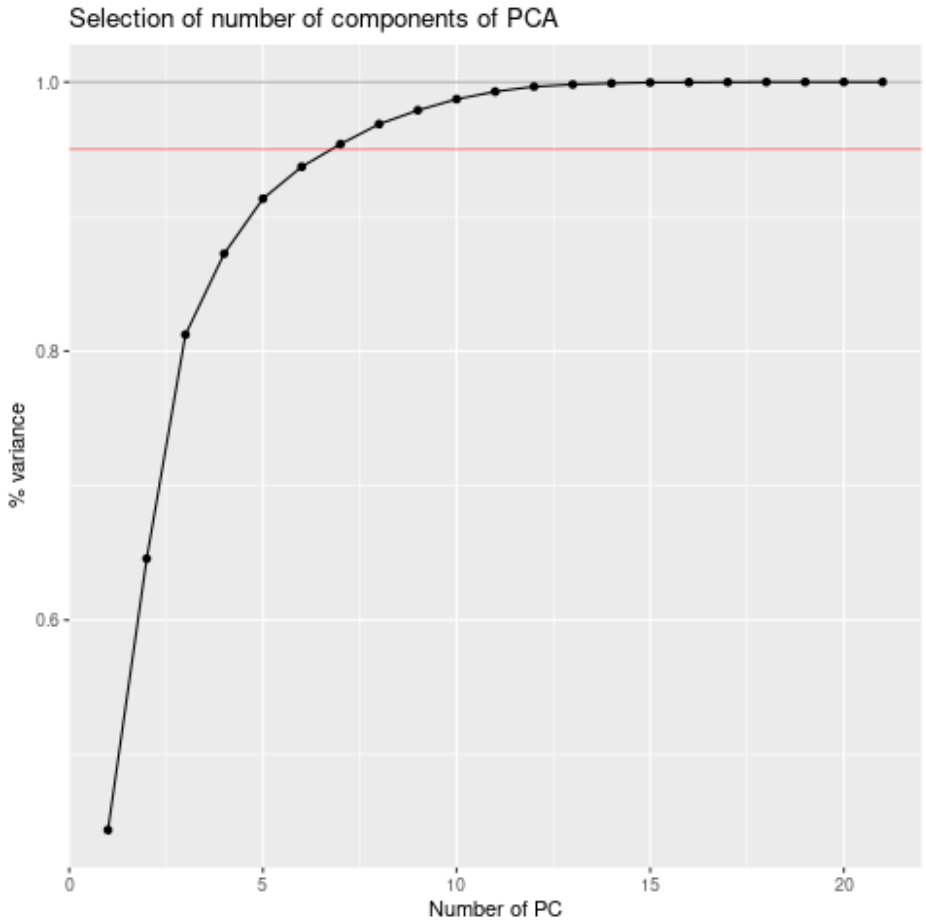

Figura 3: Plot della varianza cumulativa con le dimensioni di PCA.

accuracy	0.75
precision	0.96
recall	0.71
f_1	0.82

Tabella 5: Results on the DBScan cluster.

	HET	HOM
$CLUSTER_0$	1	8
$CLUSTER_1$	25	10

Tabella 6: Distribution of the images in the cluster.

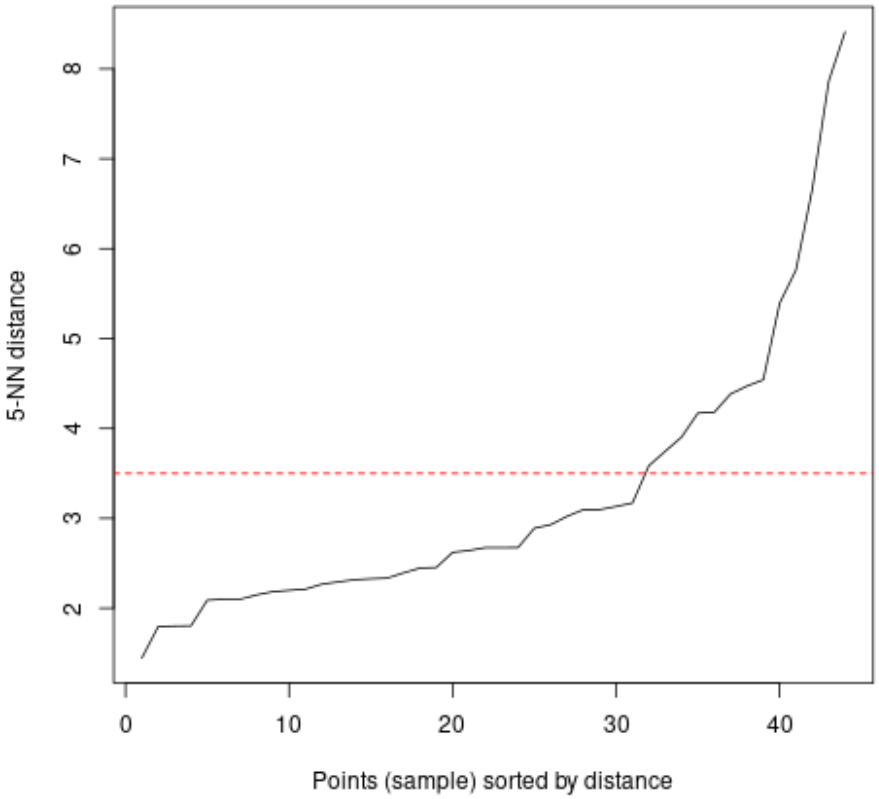

Figura 4: Scelta del eps per il DBScan

e' riuscito a trovare solo un cluster, il $CLUSTER_0$ e' un cluster costituito da prova considerate non appartenenti a nessun cluster. Quindi si tenta un altro approccio basato su un ibrido tra il clustering gerarchico e il k-means.

In questo caso pero' bisogna cercare il numero di cluster ideale, e per fare questo effettiamo il cluster per ciascun k da 2 da 15 e plottiamo le loro misure betweenss e scegliamo un k in base al plot. Nella Figura 5 scegliamo il k=4 per non "overfittare" il cluster.

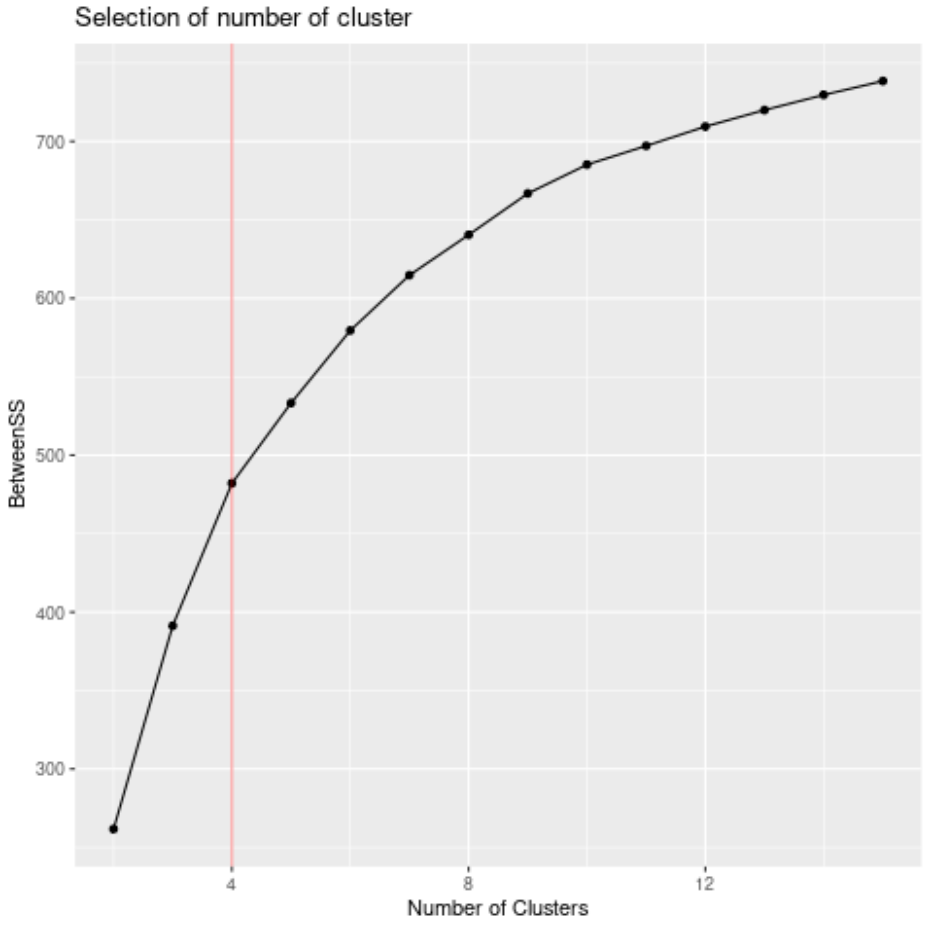

Figura 5: Plot of the BetweenSS for each k to chose the optimal one.

Nella tabella seguente si mostrano le misure ottentute dal cluster.

accuracy	0.727
precision	1
recall	0.684
f1-score	0.813

	HET	HOM
CLUSTER_1	18	8
CLUSTER_2	8	4
CLUSTER_3	0	2
CLUSTER_4	0	4

Anche il questo caso il cluster ottiene performance decenti, ma il primo modello riusciva a distinguire le due immagini in una maniera piu' decente e con meno cluster.

5 Conclusioni

Con questo Progetto si è costruito un modello statistico efficace e facilmente interpretabile da un esperto di dominio per prevedere l'eterogeneità del tumore. Si è visto che, estrapolando dall'immagine dei semplici valori indice, è possibile costruire un modello indipendente dalla dimensione dell'immagine o dalla sua risoluzione.
A livello matematico si potrebbe aumentare la prestazione del modello stimando i parametri con un numero maggiore di dati; tuttavia, in ambito medico, questo non è sempre possibile (anche perché, come espresso in [1], è possibile verificare l'eterogeneità del tumore solo in modo invasivo o con autopsia). Inoltre si potrebbero utilizzare nuove *features*, soprattutto se utili ai fini della ricerca medica.

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.