Applicazione di metodi model-based

Introduzione

Il tumore alla mammella è uno dei più diffusi, in particolare tra la popolazione femminile nella quale risulta essere il più diffuso in assoluto. In Italia, nel 2020 erano 834.200 le donne viventi con diagnosi di tumore alla mammella (fonte: *salute.gov.it*). Si tratta di un tumore che presenta una probabilità di sopravvivenza a 5 anni dalla diagnosi del 87%, per questo è importante diagnosticarlo il prima possibile.

Vista l'elevata diffusione di questo tipo di tumore e data l'importanza di una diagnosi precoce, sono presenti molti studi relativi all'argomento. Il presente report contiene i risultati dell'analisi condotta su dati relativi a tumori alla mammella, sui quali sono stati applicati metodi che permettono di classificare la massa tumorale come benigna o maligna.

Materiali e metodi

Materiali

Il data set oggetto dell'analisi "*Breast Cancer Data"* è composto dai dati relativi a 569 pazienti che presentavano tumore alla mammella. Per ogni paziente sono state rilevate 30 variabili relative ai nuclei cellulari ottenuti da un'immagine digitalizzata di un ago aspirato sottile (FNA) di una massa mammaria. A queste misurazioni sono state aggiunte altre due variabili relative all'ID del paziente e alla diagnosi del tumore, classificato come benigno o maligno.

Metodi

Inizialmente è stata svolta un'analisi esplorativa dei dati disponibili, effettuando una selezione delle variabili per poter svolgere più facilmente le analisi grafiche. In seguito, avendo a disposizione l'informazione sulla reale diagnosi relativa al tumore è stato possibile applicare sia metodi *supervised* che *unsupervised*.

Come metodo non supervisionato è stato applicato il *model-based clustering*, che permette di effettuare un raggruppamento delle unità statistiche disponibili sulla base delle variabili rilevate. L'ipotesi alla base di questa tecnica è che esistano delle sottopopolazioni nelle quali il fenomeno oggetto di interesse presenta un comportamento differente. Il fenomeno può quindi essere formalizzato utilizzando misture finite.

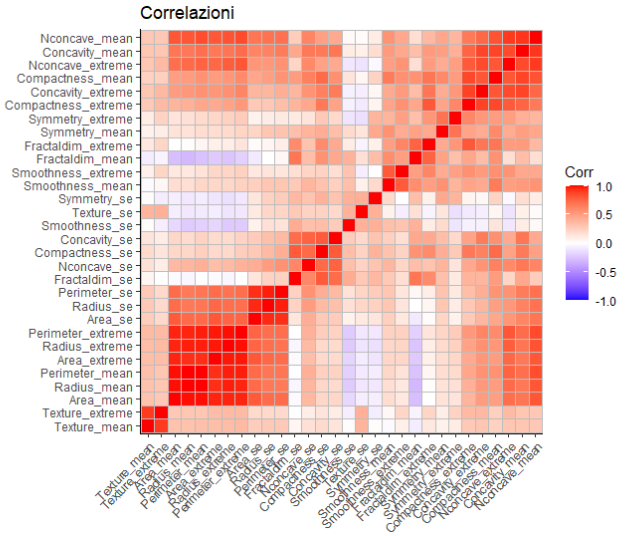
Tra i metodi supervisionati è stata scelta la *model-based classification*. Questa tecnica prevede che si istruisca un classificatore, utilizzando delle osservazioni disponibili, in modo che fornendo una nuova osservazione questa venga assegnata alla classe corretta. Tra i possibili model-based classifiers sono stati presi in considerazione:

- EDDA (Eigenvalue Decomposition Discriminant Analysis)
- MDA (Mixture Discriminant Analysis)

Risultati

Visualizzando la struttura del data set, che ha dimensioni 569×32, è stato osservato che le variabili presenti sono di natura continua ad eccezione della variabile *ID* (numerica intera) e della variabile *Diagnosis* che indica la natura del tumore, codificata nel modo seguente: *B=Benign* e *M=Malignant*. Inoltre, non sono presenti missing values.

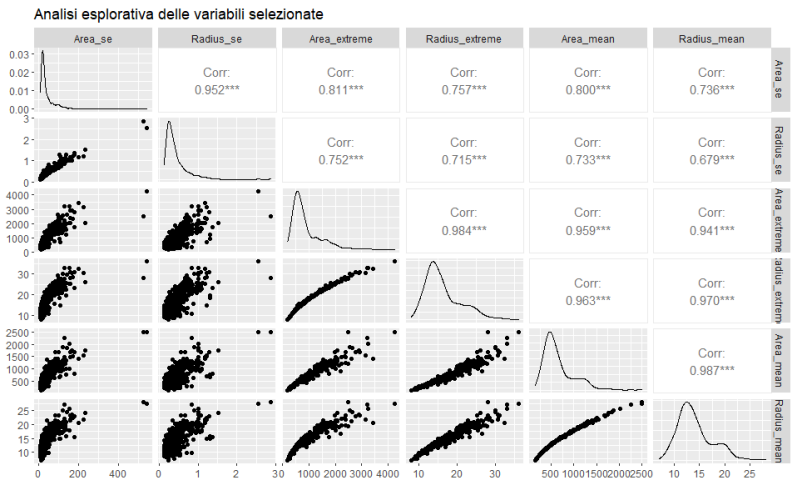
Da una prima analisi delle correlazioni, eseguita considerando tutte le variabili, emerge un'elevata correlazione tra alcune di esse. Questo è dovuto in gran parte al fatto che le variabili presenti nel dataset sono una sintesi di alcune misure rilevate su più nuclei cellulari, nel singolo paziente. Se si considera l'area dei nuclei cellulari, ad esempio, è possibile notare che nel dataset sono presenti tre variabili che sintetizzano la media, lo standard error e l'estremo per questa misura. Di conseguenza, le tre variabili Area_mean, Area_se e Area_extreme presentano elevata correlazione.



[Figura 1] Rappresentazione grafica della matrice di correlazione

Dato l'elevato numero di variabili, per poter procedere con le prime rappresentazioni grafiche, è stato necessario effettuare una selezione. Utilizzando la library `clustvarsel` e l'omonima funzione, sono state individuate le variabili fondamentali per svolgere la clusterizzazione delle osservazioni utilizzando una procedura di tipo *stepwise*, sia forward che backward. La procedura ha selezionato sei variabili: Area_se, Radius_se, Area_extreme, Radius_extreme, Area_mean e Radius_mean.

Utilizzando le variabili individuate, è stato rappresentato il grafico della [Figura 2]. In tale grafico, si nota che le distribuzioni marginali delle variabili presentano evidente asimmetria e non presentano multi-modalità. L'assenza di multi-modalità non permette di avere la certezza della presenza di gruppi considerando un numero maggiore di variabili. Tuttavia, non si è nemmeno certi della loro assenza. Osservando anche gli scatterplot bi-dimensionali è difficile individuare dei gruppi perché i punti sono molto vicini tra loro e, in alcuni casi, sono quasi perfettamente allineati sulla bisettrice. Inoltre, come già detto in precedenza, le correlazioni sono molto elevate e questo giustifica ulteriormente l'allineamento dei punti sulla bisettrice.



[Figura 2] Analisi esplorativa con scatterplot, correlazioni e densità marginali

Model-based clustering

La tecnica del model-based clustering è stata applicata ipotizzando di non conoscere il reale numero di gruppi, cioè la possibile diagnosi per il tumore (benigno o maligno), in quanto si tratta di un metodo non supervisionato.

Utilizzando la funzione `Mclust`, che esegue una stima attraverso l'algoritmo EM, è stato individuato il modello migliore per effettuare il clustering delle osservazioni. Sono state prese in considerazione diverse tipologie di modelli, imponendo vincoli di tipo geometrico sulla struttura di variabilità, e un numero variabile di componenti per la mistura.

Avendo come obiettivo principale il clustering, come criterio di valutazione della bontà dei modelli è stato utilizzato l'ICL (Integrated Complete Likelihood). In ogni caso, è stato osservato anche il BIC per poter effettuare una valutazione migliore. I migliori tre modelli che sono stati identificati sono gli stessi, sia effettuando la selezione sulla base del criterio ICL sia considerando il BIC. Nella [Tabella 1] sono riportati i valori di entrambi i criteri per i top 3 models individuati.

MODEL	ICL	BIC
VVE, 5	41221.88	41245.88
VVE, 6	41209.70	41243.18
VEE, 7	41184.41	41253.47

[Tabella 1] Top 3 models selezionati in base al criterio ICL

Il modello scelto ha volume e shape variabili, mentre l'orientation è equal (VVE), e un numero di componenti pari a cinque. Tale modello sarebbe il second best sulla base del BIC e la sua entropia risulta pari a 30 circa.

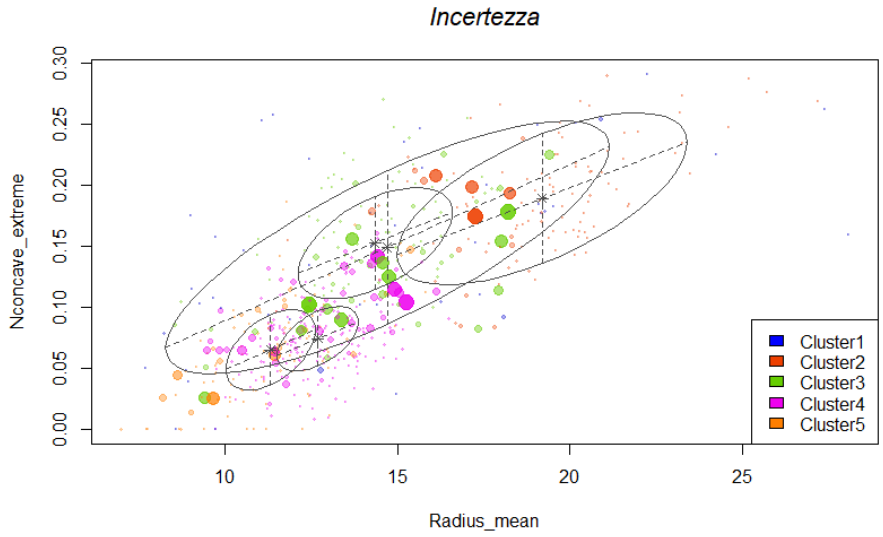
Per eseguire una valutazione della bontà del clustering effettuato adottando il modello VVE con cinque componenti, sono state calcolate alcune misure. Avendo a disposizione le vere etichette, espresse dalla variabile *Diagnosis* del dataset, è stato possibile calcolare anche CER (Classification Error Rate) e ARI (Adjusted Rand Index) oltre all'incertezza, la quale non richiede che sia nota la reale classificazione delle osservazioni. Nella [Tabella 2] sono stati riportati valori ottenuti.

CER	0.6485062
ARI	0.1348091
UNCERTAINTY	0.0522673

[Tabella 2] Misure di valutazione della bontà del clustering, modello VVE,5

La proporzione delle osservazioni misclassificate espressa dal CER è molto alta: sono state misclassificate 369 osservazioni su 569. Ciò è dovuto al fatto che è stato selezionato un modello che prevede l'esistenza di cinque clusters, mentre è noto che nella realtà sono solo due. Anche il valore dell'ARI non è soddisfacente, poiché assume un valore prossimo allo zero. Questo significa che le due partizioni confrontate sono vicine all'essere indipendenti e per questo il modello non effettua un buon clustering. L'incertezza, invece, ha un valore abbastanza basso considerando che il valore massimo che può assumere per ogni osservazione è pari a 0.8, che si ottiene qualora le stime delle *posterior probabilities* siano le stesse per ogni gruppo, cioè $1/k = 1/5 = 0.2$. Nella valutazione di questa misura, bisogna tenere conto che essa non tiene conto della reale classe di appartenenza delle osservazioni.

Dalla rappresentazione grafica dei gruppi e delle incertezze in [Figura 3] si osserva come molti punti possano essere attribuiti ad un diverso cluster e come i vari gruppi siano quasi sovrapposti tra loro e difficilmente distinguibili. Inoltre, si nota la forma dei clusters secondo il modello selezionato. Infatti, è evidente come le ellissi abbiano la stessa orientation mentre le dimensioni e le forme siano differenti tra loro.



[Figura 3] Rappresentazione grafica dei clusters e dell'incertezza

Nella [Tabella 3] sono riportati i valori delle distanze di Kullback-Leibler simmetrizzate. Dai valori ottenuti risulta evidente come alcuni gruppi siano poco distanti tra loro. In particolare, il Cluster3 risulta considerevolmente vicino al Cluster2 e al Cluster4. Al contrario, il Cluster1 risulta essere il più distante da tutti gli altri. Approfondendo i risultati del clustering, contenuti nella [Tabella 4], si osserva che tale gruppo interessa un numero esiguo di unità statistiche rispetto agli altri e la sua *mixing probability* è molto inferiore alle altre.

Cluster	1	2	3	4	5
1	0	177.678	297.322	751.816	200.625
2	177.678	0	38.944	163.982	117.639
3	297.322	38.944	0	33.904	50.430
4	751.816	163.982	33.904	0	56.048
5	200.625	117.639	50.430	56.048	0

[Tabella 3] Distanze di Kullback-Leibler

Cluster	n	Mixing prob.
1	45	0.0791
2	108	0.1899
3	104	0.1831
4	203	0.3557
5	109	0.1922

[Tabella 4] Caratteristiche dei clusters

Sapendo che in realtà in gruppi sarebbero due, provando a fornire questo parametro all'algoritmo per la scelta del modello, sono stati ottenuti risultati differenti. Il modello migliore, selezionato sulla base dell'ICL è risultato essere un VVV con due componenti. Da questo risultato, si può osservare come la riduzione (imposta) del numero di componenti abbia portato alla selezione di un modello più complesso.

CER	0.1370826
ARI	0.5253845
UNCERTAINTY	0.0008134

[Tabella 5] Misure di valutazione della bontà del clustering, modello VVV,2

La [Tabella 5] riporta le misure di bontà del clustering per questo secondo modello. È possibile osservare come un modello VVV con due componenti sia più efficiente del precedente in termini di clustering. Infatti, il CER è prossimo allo zero e questo indica un'elevata accuratezza nella clusterizzazione. Allo stesso modo, l'ARI è abbastanza alto quindi le due partizioni confrontate sono molto più simili rispetto alle precedenti; risultato spiegato dal fatto che entrambe le partizioni prevedono l'esistenza di due gruppi, mentre per il modello precedente le partizioni confrontate prevedevano una cinque clusters e quella reale solo due. Infine, si nota che l'incertezza è molto più bassa.

Model-based classification

Per poter compiere una classificazione di tipo model-based, il data set è stato diviso in *training set* e *test set* (rispettivamente 75% e 25% delle osservazioni). Trattandosi di un metodo supervisionato è stata utilizzata anche la variabile target: *Diagnosis*.

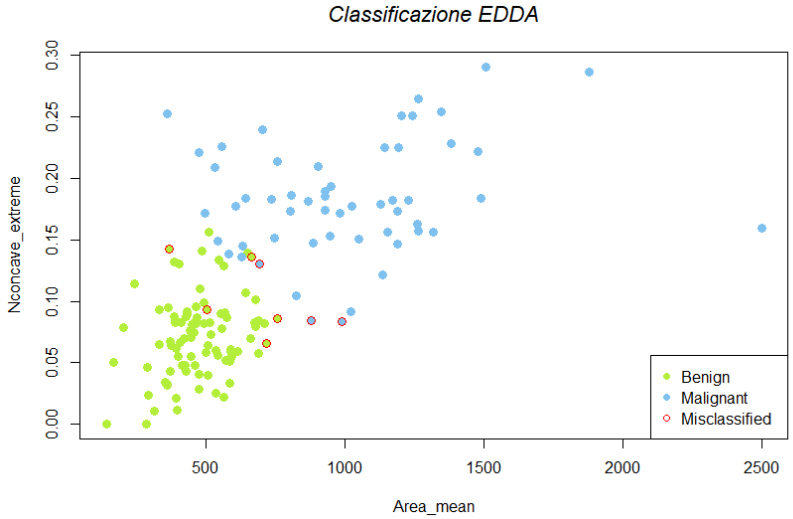
Per individuare il miglior modello per la classificazione sono stati istruiti cento classificatori utilizzando la funzione `mixmodLearn`. Nei cento classificatori sono stati individuati quattro modelli distinti per ciascuno dei quali sono stati individuati i valori minimi dei criteri CV (10-fold Cross Validation) e BIC. I risultati sono riportati nella [Tabella 6].

EDDA MODEL	VSO MODEL	CV	BIC
Gaussian_pk_Lk_Ck	VVV	0.0352	-27806.542
Gaussian_pk_L_Ck	EVV	0.0352	-26913.4592
Gaussian_pk_L_Dk_A_Dk	EEV	0.0329	-25827.1138
Gaussian_pk_L_C	EEE	0.0352	-24842.3226

[Tabella 6] Criteri CV e BIC per i model-based classifiers

Il modello che presenta il minor valore di CV è pk_L_Dk_A_Dk (EEV), mentre osservando il BIC sarebbe da preferire un modello VVV. Tuttavia, è noto che il criterio BIC è utile per valutare l'adattabilità del modello ai dati, mentre in questo caso si è più interessati al CV che costituisce una stima del MER (Misclassification Error Rate).

Dopo aver individuato il modello definitivo per la classificazione, è stato istruito il classificatore utilizzando le osservazioni presenti nel *training set*. Per valutare il classificatore, è stata eseguita la previsione della classe di appartenenza delle osservazioni presenti nel *test set*. Come si può vedere dalla [Figura 4], il risultato ottenuto si è rivelato molto soddisfacente, in quanto il MER stimato assume un valore pari a 0.0559 e le osservazioni misclassificate sono 8 su 143 del *test set*.



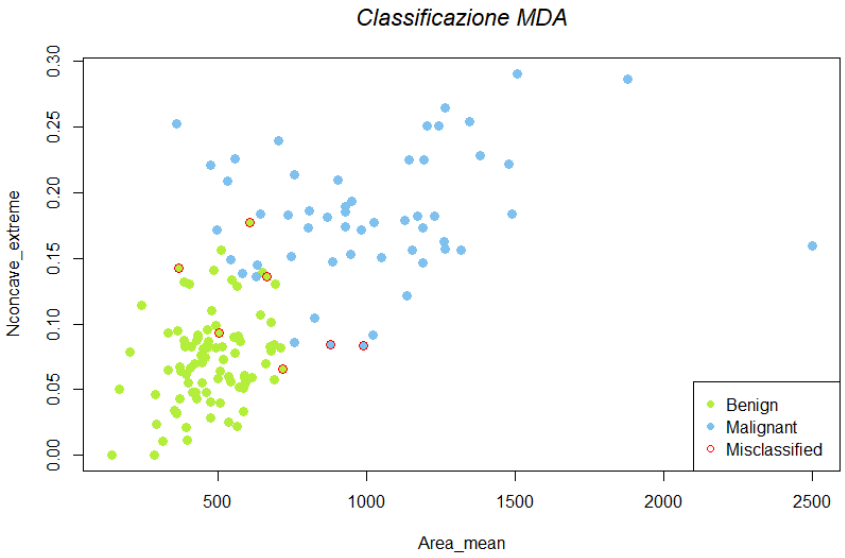
[Figura 4] Classificazione del test set secondo il classificatore pk_L_Dk_A_Dk

Per avere una visione più ampia in merito alla classificazione, sono stati presi in considerazione anche i classificatori MDA. Utilizzando la funzione `MclustDA` è stata ottenuta la stima di un modello mistura di misture che permette di avere maggiore flessibilità. Il modello individuato, riportato nella [Tabella 7], identifica per entrambi i gruppi lo stesso modello VEE, ma con diverso numero di componenti.

Classes	n	%	Model	G
B	271	63.62	VEE	3
M	155	36.38	VEE	2

[Tabella 7] Caratteristiche delle misture dei due gruppi

Tale modello risulta più complesso rispetto a quello individuato con il classificatore di tipo EDDA, ma la maggiore complessità porta ad una migliore classificazione, come mostra la stima del MER che risulta pari a 0.0489. Nella [Figura 5] si può osservare la classificazione effettuata dal classificatore MDA e le unità statistiche misclassificate.



[Figura 5] Classificazione del test set secondo il classificatore MDA

Conclusioni

Le analisi svolte hanno mostrato come il *model-based clustering* non sia efficiente per questo dataset. Infatti, i dati disponibili non permettono una chiara identificazione dei due clusters reali a causa della vicinanza e similitudine tra le osservazioni. Per individuare correttamente i due gruppi è stato necessario imporre il numero di componenti desiderato, ma il risultato ottenuto ha portato ad un modello con la massima complessità possibile. Il modello selezionato dall'algoritmo, senza imporre il numero di classi, si è rivelato più semplice ma con un numero di componenti maggiore. Questo evidenzia il trade-off tra un modello più semplice, che porta ad un numero maggiore di gruppi, e un numero inferiore di gruppi, che al contrario richiede un modello molto articolato. Nel caso esaminato, tuttavia, con un numero maggiore di componenti la semplificazione del modello è risultata lieve.

Per quanto riguarda la *model-based classification*, da una valutazione complessiva, emerge che per il data set in oggetto un classificatore di tipo EDDA risulta adatto, in quanto la maggior complessità del MDA classifier non è compensata da un MER significativamente migliore. Si nota, infatti, che il MDA classifier classifica correttamente solo un'unità statistica in più rispetto all'EDDA classifier, pertanto non si ritiene necessario adottare il modello più complesso.