Classificazione Asteroidi Nasa

Lavoro di : Matteo Baldanza

April 30, 2020

Gli asteroidi sono corpi celesti formati da grossi blocchi di roccia o di roccia e metallo, che gravitano nel nostro sistema solare.Gli scienziati ritengono che rappresentino i relitti relativi alla formazione del Sistema Solare, avvenuta 5.000 milioni di anni fa. Un asteroide è un corpo roccioso dalla forma irregolare il cui diametro può variare da qualche decina di metri a migliai di km. L'importanza di osservarli e di studiarli è dovuta al fatto che ne esistono alcuni che hanno un orbità che si interseca con quella della terra e un asteoride di mini dimensioni (qualche km) provocherebbe disastri inimmaggiabili se si scontrasse con il nostro pianeta(basti pensare all'estinzione dei dinosauri). Da questo la necessità di controllarli al fine di capire le possibilità che diventino pericolosi per la nostra civilità

Intoduzione

Il lavoro è eleborato fruendo un dataset della Nasa [1] costituito da 4687 osservazioni e 40 variaibili. Ogni unità statistica rappresenta un asteroide , le variabili rappresentano le sue caratteristche eccetto una che indica come l'asteroide sia stato classificato dall'agenzia spaziale : "Potenzialmente pericoloso(*True*) o potenzialmente non pericoloso(*False*)". L'obbiettivo del lavoro è di trovare tramite l'utilizzo di misture il miglior modello di classificazione. Lo scopo della classificazione è di predire la classe di appertenza,in questo caso asteroide pericoloso o no, di un nuovo tipo di unità statistica.

1 Riduzione della dimensionalità

1.1 Le prime osservazioni

Il dataset è costituito da 40 variabili per cui c'è bisogno,prima di applicare i vari algoritmi,di ridurne il numero ,cercando di capire quale di queste possano essere più significative al fine dell'analisi. La primissima osservazione è che variabili come l'id,il nome e le date di osservazione degli asteroidi sono irrilevanti e per questo non avranno mai nessun tipo di impatto statistico importante nell'analisi.
Per quanto riguarda i valori mancanti, in questo caso si è fortunati perchè il dataset ne è completamente privo. Prima di effettuare delle vere e proprie analisi sulla scelta delle variabili bisogna,come sempre ,visionare la numerosità delle etichette per capire come sono distribuite :

False	True
3932	755

Si nota che le classi sono sbilanciate, per cui si potrebbero adottare tecniche per il ribilanciamento, che in questo lavoro non vengono utilizzate, per migliorare le performance dei vari algoritmi. Si procede quindi con un'analisi delle correlazioni. A tal proposito si ricorda che alte correlazioni fra variabili sono indice quasi sempre di medesime spiegazioni dovute da una forte dipendenza delle due. Il risultato ottenuto è il seguente :

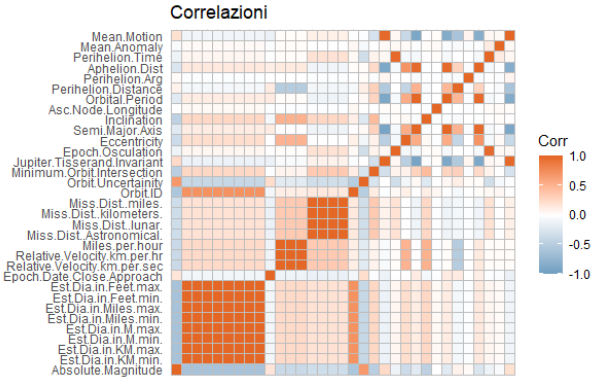

Figure 1: *Correlazioni fra tutte le colonne*

Si vede dal grafico che alcune variabili sono molto corre-

[1] https://www.kaggle.com/shrutimehta/nasa-asteroids-classification

late e andando ad indagare sui nomi si scopre che l'alta correlazione è dovuta semplicemente al fatto che molte colonne esprimono in realtà la stessa cosa. Ad esempio le colonne "*Est.Dia.in.KM.max.,Est.Dia.in.KM.min.,*
Est.Dia.in.M.max.,Est.Dia.in.M.min.,
Est.Dia.in.Miles.max.,Est.Dia.in.Miles.min.,
Est.Dia.in.Feet.max.,Est.Dia.in.Feet.min." ,indicano tutte la dimensione dell'asteroide ma in differenti unità di misura e quindi non ha nessun senso portarsele avanti. Si è deciso a tal fine di mantenere solo le variabili espresse in km. Lo stesso identico discorso vale per le seguenti variabili che esprimono la distanza minima asteroidi/terra : "*Miss.Dist..Astronomical.,Miss.Dist..lunar.,*
Miss.Dist..kilometers.,Miss.Dist..miles.".Anche in questo caso si è tenuta la sola variabile con i km. Per ultimo si possono osservare tali variabili che esprimono la velocità dell'asteoride : "*Relative.Velocity.km.per.sec,*
Relative.Velocity.km.per.hr,Miles.per.hour" ,e ancora una volta si è deciso di mantenere quella con i km.
A questo punto rimangono 23 variabili.

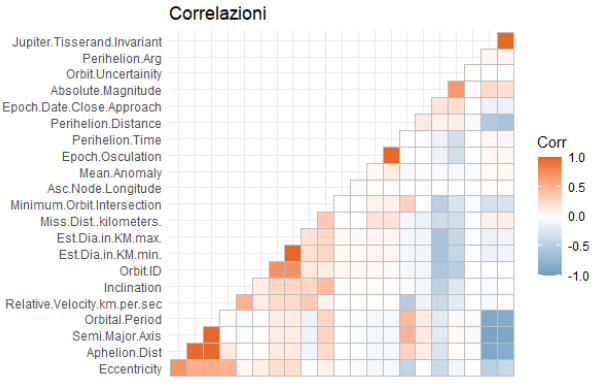

Figure 2: *Correlazione tra le variabli rimaste*

Sebben si possano togliere variabili con correlazione alta si è deciso di tenerne alcune in quanto ritenute importanti per la spiegazione delle etichette. Si mantengono le variabili come ad esempio "*Aphelion.Dist*" [2], "*Orbital.Period*" [3] e *Epoch.Osculation*" [4] mentre si tolgono la variabili "*Est.Dia.in.KM.min.*" *e* "*Est.Dia.in.KM.max.*" in quanto la misura di un asteroide è un informazione che potrebbe benissimo essere rappresentata da un'altra variabile. Questa misura è infatti correlata,ma negativamente, con la variabile *Absolute.Magnitude*[5] il che indicherebbe che più un asteroide è grande meno magnitudine/luce ha,il che porterebbe ad un assurdo. Occorre notare però che la scala delle magnitudini è inversa e quindi un oggetto molto luminoso ha un valore inferiore rispetto ad uno meno luminoso. Il dataset risulta ora composto da 20 variabili che sono ancora un numero troppo elevato.

1.2 Riduzione dimensionalità tramite PCA

Un approccio di riduzione della dimensionalità è quello della PCA(principal component analysis). Lo scopo della tecnica è quello di ridurre il numero più o meno elevato di variabili che descrivono un insieme di dati a un numero minore di variabili latenti, limitando il più possibile la perdita di informazioni. Tale nuove variabili sono le componenti principali formate da un peso dato da ogni variabile.
Una tecnica è quella di costruirsi la matrice dei punteggi , che sarà formata dalle unità statistiche del dataset che avranno un determinato punteggio in base al peso di ogni variabile sulle componente principale. In questa maniera seppur efficace si perde l'interpetabilità delle variabili per cui non si approfondirà tale strategia.
Una seconda possibilità è la seguente. Si considerano le prime n componenti principali che spiegano una certa variabilità, ad esempio l'80% ,e se ne trovano le variabili che ne danno il maggior peso. Utilizzando questa tecnica sono state selezionate le seguenti modalità : "*Semi.Major.Axis ,Absolute.Magnitude , Perihelion.Time,Perihelion.Distance*
Orbit.ID ,Perihelion.Arg".La rappresentazione grafica degli incroci di queste variabili non ha purtroppo prodotto nessun cluster evidente e data anche la perfomance non buona che si vedrà più avanti si è scelto di provare anche un metodo differente per la riduzione della dimensionalità

1.3 Riduzione dimensionalità tramite GLM

Quello che si vuole capire all'interno del dataset è quali variabili spiegano meglio la suddivisione delle classi. Per questo proposito l'idea è di utilizzare un glm(modello lineare generalizzato) che mi riesca a spiegare una variabile risposta di tipo dicotomico/categorica in funzione del resto di variabili. Un modello noto è quello appartenente alla famiglia binomiale. In sostanza, la regressione logistica permette di generare un risultato che, di fatto, rappresenta una probabilità che un dato valore di ingresso appartenga a una determinata classe. Sfruttando questa conoscenza e applicando un algoritmo di tipo stepwise backward si riescono a identificare quali sono le variabili significative all'interno del modello,cioè si comprendono quali sono quelle che riescono a spiegarmi meglio la variabile dipendente. L'algoritmo stepwise consiste in una regressione graduale in cui la scelta delle variabili predittive viene effettuata mediante una procedura automatica. In ogni passaggio, viene considerata una variabile per l'aggiunta(forward) o la sottrazione(backward)

[2] Il punto più lontano di un'orbita dal sole
[3] Il tempo impiegato da un corpo in orbita per compiere una rivoluzione completa attorno al Sole
[4] E' l'orbita gravitazionale di Keplero (cioè un'ellittica o altra conica) che un corpo avrebbe intorno al suo centro se le perturbazioni fossero assenti
[5] E' una misura della luminosità intrinseca di un oggetto senza tener conto delle sue variazioni di luminosità

dall'insieme di variabili esplicative in base a un criterio specifico($F\,o\,T$).Utilizzando tale procedura si è trovato tale risultato :

```
Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  4.872e+02  2.154e+02   2.262 0.023688 *
Absolute.Magnitude          -1.669e+00  9.349e-02 -17.854  < 2e-16 ***
Orbit.ID                    -2.123e-02  2.777e-03  -7.646 2.08e-14 ***
Orbit.Uncertainity          -2.729e-01  3.876e-02  -7.042 1.90e-12 ***
Minimum.Orbit.Intersection  -1.202e+02  5.880e+00 -20.439  < 2e-16 ***
Jupiter.Tisserand.Invariant  6.841e+00  2.062e+00   3.319 0.000905 ***
Semi.Major.Axis              2.848e+00  1.049e+00   2.715 0.006627 **
Inclination                  3.945e-02  1.415e-02   2.788 0.005305 **
Perihelion.Distance         -2.279e+00  1.028e+00  -2.217 0.026629 *
Perihelion.Time             -1.905e-04  8.765e-05  -2.174 0.029715 *
Mean.Anomaly                 1.083e-03  7.508e-04   1.442 0.149209
Mean.Motion                 -2.189e+01  6.477e+00  -3.380 0.000725 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 3: *Risultato del modello*

Sebben il metodo di minimizzazione dell'AIC prediliga l'inserimento della variabile *Mean.Anomaly*,dato che risulta non significativa si decide di toglierla. Queste 10 variabili definitivamente scelte sono le più importanti all'interno del dataset e sono quelle che ottimizzeranno i risultati degli algoritmi. Il grafico sottostante conferma che la scelta effettuata sia migliore rispetto a quella tramite pca.

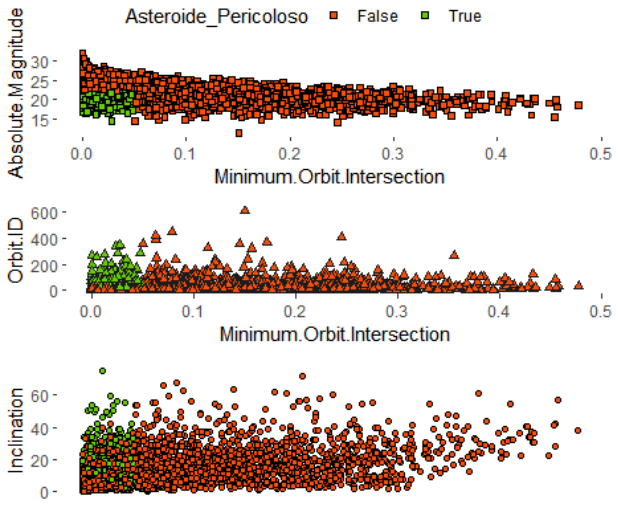

Figure 4: *Cluster evidenti graficamente*

I cluster sono evidenti. Appare anche ovvio da tale grafico che gli asteroidi classificati con "*Minimum.Orbit.Intersection*" [6] sopra ad una certa soglia non possano essere considerati pericolosi mentre quelli al di sotto non è detto che lo siano ma dipende da differenti valori assunti dalle variabili. Con questi risultati si ritiene che il metodo migliore di selezione sia quest'ultimo.

[6] È definita come la distanza tra i punti più vicini delle *Epoch.Osculation*(nota 4) di due corpi

2 Classificazione con modelli EDDA

I modelli EDDA sono modelli mistura che sfruttano il calcolo di una probabilità a posteriori per assegnare ad una determinata classe un'unità statistica.In questo lavoro si tratteranno soltanto esempi di misture con distribuzioni normali [7]. Il loro vantaggio rispetto alle tecniche di classificazione come *LDA e QDA* sta nel fatto di non dare nessun vincolo alle matrici di covarianze. In questi modelli infatti si scompongono così :

$$\underset{d\times d}{\Sigma_\kappa} = \underset{1\times 1}{\lambda_\kappa} \times \underset{d\times d}{D_\kappa} \times \underset{d\times d}{A_\kappa} \times \underset{d\times d}{D'_\kappa} \qquad (1)$$

Dove κ ,k-esima componente di ogni mistura, λ autovettore che deriva da $\tilde{A} = \lambda \times A$ e ne indica il volume, D matrice di autovettori normalizzati che ne indica l'orientamento in R^d e A matrice diagonale degli autovettori divisi per λ che ne indica la forma.Tutte queste caratteristiche modellano la funzione di densità di ogni componente. Da tale scomposizione in base alle scelte delle matrici si possono formare differenti tipi di modelli, ad esempio con "EEE" ,che rappresenta l'opzione più semplice, si indica volume uguale(λ uguale per ogni componente),stesso orientamento(D_k identiche) e stessa forma(A uguali). La probabilità a posteriore,il peso di ogni componente nel modello,i parametri delle distribuzioni normali e la numerosità di ogni gruppo vengono trovate tramite algoritmo **EM(Expectation-Maximisation)**.

2.1 Scelta del Test-Set

Il training set è un insieme di dati che vengono utilizzati per addestrare un sistema supervisionato(etichette note).Consiste in un vettore di input a cui è associata una risposta o una determinata classificazione. Una volta eseguito, l'algoritmo(in questo caso EM) apprende, in base alla risposta o alla classificazione, quali caratteristiche "discriminano" gli elementi appartenenti alle differenti categorie.Una volta effettuata questa fase, la correttezza dell'algoritmo viene verificata eseguendo lo stesso modello del training sul test set. Il test-set è un insieme di unità statistiche con le stesse modalità del training-set ,ma con dati che non vengono utilizzati durante la fase di scelta del modello, allo scopo di capire se si è fatto un buon lavoro o meno. In questo caso si è deciso,in maniera del tutto arbitraria di separare il dataset mantendendo un training con l'80% delle osservazione e quindi un test-set con il restante 20%(937 osservazioni)

2.2 EDDA con PCA

Per non scartare il risultato ottenuto con PCA , si utilizzano le variabili scelte in 1.2 per l'utilizzo

[7] Si assume verificata tale ipotesi in questo lavoro

dell'algoritmo. La scelta del modello mistura che spiega la suddivisione di ogni componente e che performi al meglio nella classificazione è scelto tramite il criterio k-fold-cross-validation. Il metodo consiste nella suddivisione del training-set in k gruppi scelti a priori con una certa logica,pochi gruppi porteranno a varianza bassa ma bias alto e viceversa.I gruppi uno alla volta costituiranno il test-set e alla fine si avrà un indice di nome MER che ci dirà la percentuale di miss-classification.C'è quindi bisogno di un trade-off tra i due. Il modello scelto sarà quello che lo minimizza. In questo caso si è creato un ciclo che itera più volte l'algoritmo EM e per ogni iterazione si considera il valore del CV più piccolo. Il risultato è stato il seguente :

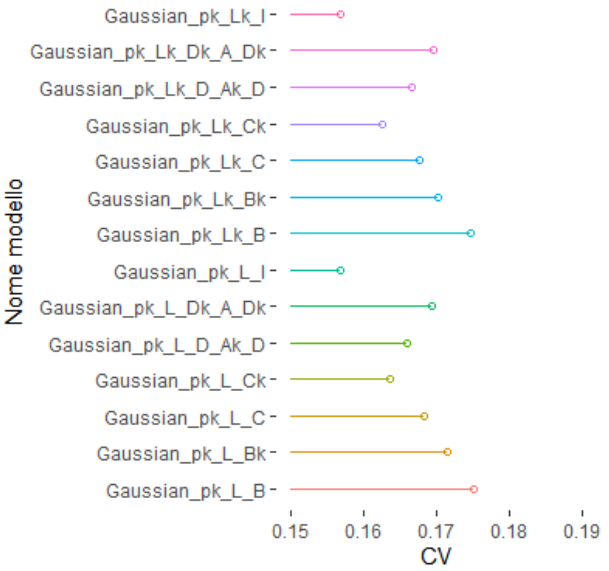

Figure 5: *K-Fold-Cross-Validation*

Il miglior modello è un modello P_k, L, I ovvero un "EII" ove I matrice d'identità. Ora che si è scelto il modello migliore nel training set quello che si fa è di prendere il test-set,ripetere l'algoritmo su tali unità statistiche e osservare se tali unità vengono classificate correttamente.Chiaramente il modello utilizzato sarà quello che ha minimizzato il CV in precedenza.

$$Accuracy = 0.8217716$$

Il risultato ci dice che l'82% degli asteroidi appartenti al test-set è stato classificato correttamente come potenzialmente pericoloso/non pericoloso.Il problema dell'algoritmo in questo caso è che nessuna osservazione viene classificata come "True", ovvero l'accuracy più alta si ottiene assegnando tutte le osservazioni alla classe "False" [8]. E' presente quindi un alto livello di errori di prima specie dovuto al fatto che il modello più semplice(quello che assegna tutte le osservazioni ad una etichetta) ha accuracy maggiore! Con tale risultato si ha che molti asteroidi vengono classificati come pericolosi ma in realtà non lo sono,questi sono i cosidetti falsi-positivi. Un classificatore di questo tipo non è un risultato significativo per quest'analisi.

[8] Si fa notare che tale errore è dovuto allo sbilanciamento delle classi nel training-set

2.3 EDDA con variabil GLM

Dato che il risultato ottenuto con PCA non è soddisfacente si è provato ad utilizzare le variabili scelte nel paragrafo 1.3. Il procedimento è analogo a quello effettuato in precedenza e senza rispiegare il tutto si procede ad una visualizzazione del risultato del CV :

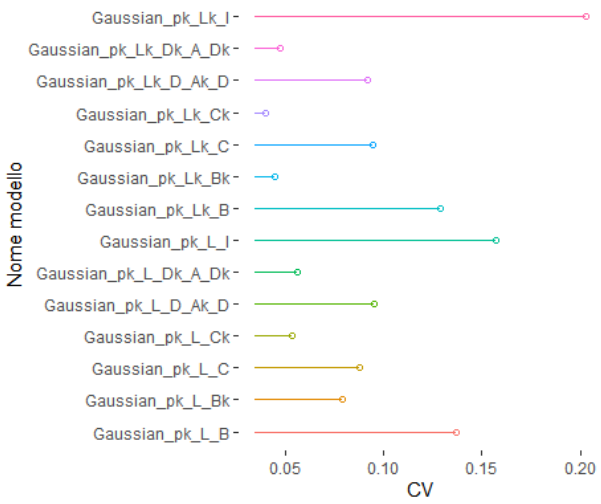

Figure 6: *K-Fold-Cross-Validation*

Il miglior modello,in questo caso,risulta P_k, L_k, C_k ovvero un "VVV". Per quanto riguarda invece le performance si hanno dei notevoli miglioramenti rispetto a prima :

		Teorica	
		False	True
Predetti	False	751	19
	True	22	145

Accuracy : 0.9562
95% CI : (0.9411, 0.9684)
Sensitivity : 0.9715
Specificity : 0.8841

L'errore di prima specie è sicuramente,molto più basso rispetto al precedente caso e quindi si può dire che effettivamente la scelta delle variabili effettuata con GLM sia stata ottima per questo tipo di lavoro. In questo caso ogni asteoroide verrà classificato correttamente dall'algortimo con una percentuale del 96%!

3 Classificazione con MDA

Sebben il risultato sia gia buono, si è testato anche un metodo di classificazione di tipo MDA. Tale metodo è simile a quello precedente ma con la differenza che ogni componente del modello mistura non è una semplice variabile casuale ma è a sua volta una mistura formata da κ componenti. Per riassumere il modello è descritto da una mistura formata da componenti che sono esse stesse delle misture! In questo caso l'algoritmo sarà sicuramente più pesante dal punto di vista computazionale ma non si può escludere che tale metodo sia migliore,anche perchè questo non vincola le due componenti ad essere distribuite come normali. La scelta dei vari modelli viene effettuata tramite minimizzazione del BIC. Questa volta non si procederà tramite le variabili scelte con PCA in quanto si è già visto che performano male.

3.1 MDA con variabili GLM

Come per EDDA per la scelta del modello migliore è bene far iterare l'algoritmo più volte per evitare errori. Il risultato finale a cui si è giunti è che le misture delle componenti hanno diversi numeri di componenti ovvero 5 e 3.Per entrambi i modelli inoltre la scelta migliore è stata un modello di tipo "VVV".A questo punto se si osserva la confusion matrix appare che, questo tipo di modello ,sia seppur di poco più performante rispetto a quanto ottenuto con i modelli di tipo EDDA.

		Teorica	
		False	True
Predetti	False	760	26
	True	10	141

Accuracy : 0.9616
95% CI : (0.9472, 0.9729)
Sensitivity : 0.9870
Specificity : 0.8443

4 Ultime osservazioni

Nell'ultima parte si vuole mostrare come per un buon modello di classificazione non è importante la numerosità delle variabili utilizzate ma bensì la qualità con cui le variabili esprimono la suddivisione dei cluster. Ad esempio in questo lavoro come è riportato nella *Figure 4* le caratteristiche che spiegano meglio le classi sono formate da queste tre differenti variabili : "*Minum.Orbit.Intersection,Absolute.Magnitude e Inclination* [9] ". Provando ad utilizzare quest'ultime all'interno di modelli di misture di tipo EDDA si ha un miglioramento :

[9] Angolo tra il piano dell'orbita e il piano dell'eclittica. Si fa notare che non è stata consideratala la variabile Orbit.ID

		Teorica	
		False	True
Predetti	False	764	6
	True	20	147

Accuracy : 0.9723
95% CI : (0.9596, 0.9818)
Sensitivity : 0.9745
Specificity : 0.9708

Per quanto riguarda invece l'utilizzo di un modello MDA si ottiene ancora una volta una perfomance maggiore , con differenti composizioni delle componenti:

Classes	Model	K
False	VVV	5
True	VVI	4

		Teorica	
		False	True
Predetti	False	770	18
	True	0	149

Accuracy : 0.9808
95% CI : (0.9698, 0.9886)
Sensitivity : 1.0000
Specificity : 0.8922

E' vero che i risultati sono vicini ma MDA performa meglio!

Conclusioni

Si è riusciti a classificare gli asteoroidi con una certezza del 98% . Per quanto riguarda la scelta finale del modello,EDDA o MDA,bisogna osservare che nessuno dei due è migliore dell'altro. Se si vuole un classificatore che prediliga la sensitivity si ricorrerà a MDA e si avrà in questo modo lo 0% di fare errori di prima specie. Viceversa se si preferisce ,invece, avere meno errori di seconda specie si opterà per un modello EDDA.