Classificatore binario per l'essenzialità dei geni in E.Coli e S.Cerevisiae

Tommaso Della Rosa
Università degli Studi di Firenze
Email:tommaso.dellarosa@stud.unifi.it

***Sommario*—L'obbiettivo dell'esercizio è quello di riprodurre lo studio eseguito Adam M Gustafson, Evan S Snitkin, Stephen CJ Parker, Charles DeLisi e Simon Kasif [1] che consiste nell'utilizzo dell' algoritmo di classificazione Bernoulli naive Bayes per l'identificazione dell'essenzialità dei geni di E.Coli e S.Cerevisiae**

1. Introduzione

L' esercizio è stato suddiviso in 4 fasi:

- Manipolazione dei Dataset
- Classificazione con Bernoulli naive Bayes
- Calcolo curve ROC(Receiver Operating Characteristic) e AUC(Area Under Curve)
- Calcolo PPV

L'ambiente di sviluppo utilizzato è Google Colaboratory, scelto per la sua velocità di esecuzione.

2. Manipolazione Dataset

Per poter classificare i dati, formati dalle caratteristiche proteiche e genomiche dei geni presenti in E.Coli e S.Cerevisiae, è stato necessario innanzitutto convertirli in numeri binari tramite la seguente regola: se il valore è minore di una certa soglia, che varia in ogni caratteristica, allora viene posto a 0, altrimenti viene posto a 1.
Per creare Trainset e Testset vengono selezionati metà dei geni essenziali e metà dei non essenziali in maniera randomica; questa operazione verrà eseguita per 100 volte così come la classificazione, in modo tale da risultare più precisa.
Dopodiché ogni Dataset viene suddiviso in subset (3 per S.Cerevisiae, 2 per E.Coli), dove le caratteristiche vengono selezionate tramite la tecninca CMIM(Conditional Mutual Information Maximization) [2], la quale seleziona le caratteristiche migliori e ne stila una classifica in base alle informazioni che ognuna di esse apporta all'essenzialità del gene; questo calcolo viene effettuato tramite la probabilità condizionata.

I subset che vengono presi in considerazione sono:

- `SC_All`
- `SC_GenProt`
- `SC_GenProt_No`
- `EC_GenProt`
- `EC_GenProt_No`

2.1. `SC_All`

Questo subset contiene le 42 migliori caratteristiche di S.Cerevisiae che possono essere ottenute facilmente dalla sequenza genomica e dopo studi intensivi sulle proteine.

2.2. `SC_GenProt`

Questo subset contiene le 16 migliori caratteristiche di S.Cerevisiae senza tenere conto degli studi sulle proteine, quindi è composto unicamente dalle caratteristiche ottenibili dalla sequenza genomica.

2.3. `SC_GenProt_No`

Contiene le stesse caratteristiche di `SC_GenProt` ma non tiene conto della phyletic retention, ossia il numero di geni rimasti perlopiù immutati durante l'evoluzione del batterio nella storia.

2.4. `EC_GenProt`

Come `SC_GenProt` con la differenza che il batterio in questione è E.Coli. Composto da 28 caratteritiche

2.5. `EC_GenProt_No`

Contiene le stesse caratteristiche di `EC_GenProt` ma non tiene conto della phyletic retention.

3. Bernoulli naive Bayes

Fa uso della probabilità condizionata, in particolare del teorema di Bayes, $p(R_k|x)=\frac{p(R_k)p(x|R_k)}{p(x)}$ dove x è il dato in input e R è la classe dei k possibili risultati. In breve la funzione classificatrice in naive Bayes è la seguente:

$$\hat{y} = \underset{k\in\{k_1\cdots k_n\}}{\operatorname{argmax}}\ p(R_k)\prod_{i=1}^{n} p(x_i|R_k) \tag{1}$$

In Bernoulli naive Bayes $p(x_i|R_k)$ è dato da:

$$p(x_i|R_k) = \prod_{i=1}^{n} p_{ki}^{x_i}(1-p_{ki})^{(1-x_i)} \tag{2}$$

Per eseguire Bernoulli naive Bayes è stata implementata la funzione `BernoulliNB()` [3], in particolare sono state utilizzate le funzioni `fit()` per il training del classificatore, `predict()` per le predizioni del test set, `predict_proba()` per le probabilità di ogni singolo gene.

4. ROC, AUC e PPV

Le ROC(Receiver Operator Characteristic) sono delle curve, che rappresentano le prestazioni di un classificatore. In particolare più la curva si avvicina all'angolo in alto a sinistra del grafico, ovvero più grande è l'area sottesa dalla curva(AUC), più il classificatore è preciso. In questo caso con il termine precisione si indica la probabilità che una predizione positiva sia effettivamente giusta.
TPR e FPR sono i parametri per il calcolo delle curve ROC e sono così definiti:

$$TPR = \frac{\mathrm{TP}}{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3}$$

$$FPR = \frac{\mathrm{FP}}{N} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{4}$$

e rappresentano, rispettivamente, il rapporto delle predizioni positive esatte(TP) con il numero totale di positivi(P) e delle predizioni positive errate(FP) con il numero totale di negativi(N).
PPV(Positive Predictive Values) invece è così definito:

$$PPV = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{5}$$

e rappresenta il rapporto del numero di predizioni positive esatte con il numero di predizioni positive totali(TP+FP). Per calcolare TPR, FPR, e AUC sono state usate, rispettivamente, le funzioni $metrics.roc\_curve()$ [4] e $metrics.roc\_auc\_score()$ [5].

5. Risultati Sperimentali

I risultati ottenuti, per quanto riguarda curve ROC e le percentuali di PPV, sono dati dai seguenti grafici.

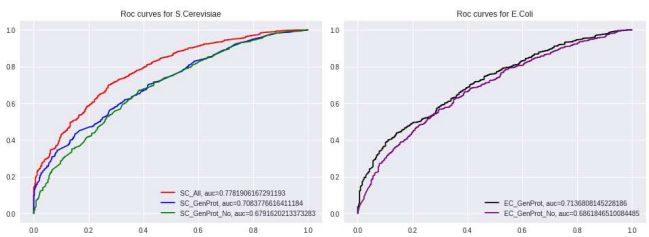

Figura 1: Curve ROC

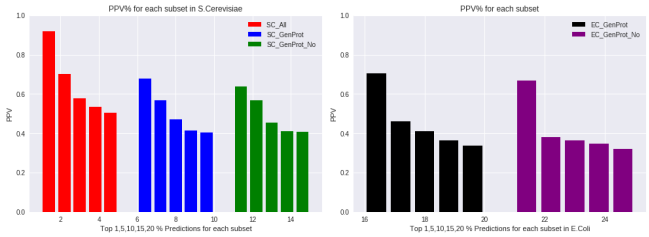

Figura 2: PPV

Dato che la classificazione è stata eseguita per 100 volte, è stato necessario calcolare la media di AUC e di PPV ad ogni iterazione per poterli rappresentare, mentre le il grafico delle curve ROC è dato dai soli risultati dell'ultima iterazione.

6. Conclusioni

Come è possibile notare in Figura 1 e 2, In S.cerevisiae il subset che ha performance migliori è `SC_All` con un valore medio di AUC pari a 0.77, a discapito dei 0.71 e 0.68 di `SC_GenProt` e `SC_GenProt_No` rispettivamente. In E.Coli il subset migliore è `EC_GenProt` con un valore medio di AUC di 0.72, mentre `EC_GenProt_No` ha 0.69. Anche per quanto riguarda i PPV si riscontra lo stesso andamento, infatti `SC_All` ha un percentuale media di PPV di circa 92% mentre `SC_GenProt` ha il 68% e `SC_GenProt_No` il 64%, così come `EC_GenProt` ha 70%, mentre `EC_GenProt_No` ha 67%.
Questi risultati erano facilmente intuibili, in quanto i subset che sono risultati i più precisi sono anche quelli che contenevano più caratteristiche. In particolare `SC_Alll` contiene anche le informazioni ottenute tramite studi in laboratorio, le quali danno un apporto significativo alla classificazione.

Riferimenti bibliografici

[1] https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-7-265

[2] https://static-content.springer.com/esm/art%3A10.1186%2F1471-2164-7-265/MediaObjects/12864_2006_648_MOESM1_ESM.xls

[3] https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.BernoulliNB.html

[4] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_curve.html

[5] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html