RELAZIONE ELABORATO

Pietro Longinetti

Naive Bayes per l'essenzialità dei geni

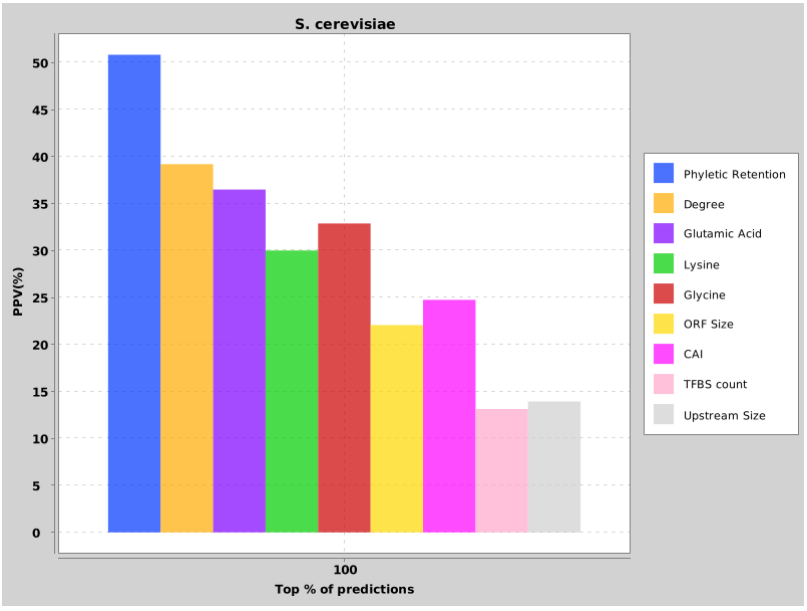
Introduzione

L'elaborato in questione fa riferimento ad un articolo di ricerca scientifica che discute sull'essenzialità dei geni: si vuole infatti predire, avendo a monte innumerevoli dati scientifici di supporto che caratterizzano la composizione e il comportamento di un gene, se un dato frammento genetico è essenziale, cioè indispensabile per la sopravvivenza per di un certo organismo, o no.
La suddetta predizione viene fatta, sotto certe ipotesi, con un classificatore binario di *Bayes* ingenuo.

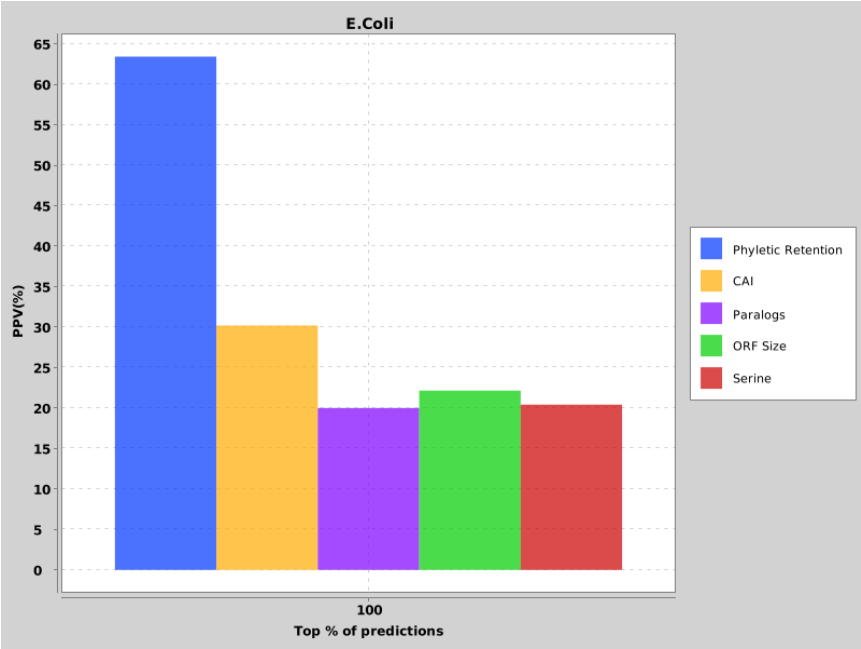
Svolgimento

Il termine "ingenuo" non sta tanto a significare che l'algoritmo di classificazione non è attento, ma che considera ogni attributo come se fosse condizionalmente indipendente dagli altri. Quindi i nostri attributi per il gene: dimensione del frammento, conteggio dei *paraloghi*, tasso di ricombinazione, grado di interazione proteica, localizzazione delle protein, ecc… saranno considerati tutti separatamente e con stesso peso predittivo. Particolarmente interessante è la caratteristica di Ritenzione Filetica, una caratteristica genomica che, misurando il numero di organismi in cui è presente un *ortologo* (un gene che durante l'evoluzione si è separato da un antenato comune ed è rimasto uguale nella sequenza genomica), rappresenta la caratteristica di gran lunga più predittiva per l'essenzialità rispetto a tutte le altre, e ciò si noterà anche sperimentalmente.

Project 4A. La prima cosa che è stata richiesta di fare per l'elaborato è stata quella di mostrare quanto vale la misura di PPV (*Positive Predicted Values*) altresì chiamata *Precision,* per ogni singola caratteristica genomica di un set di geni di *S.Cerevisiae*. Il PPV è una misura che si ottiene applicando il classificatore *Naive Bayes* ad un'istanza di dati, utilizzandone una parte come *training* e una parte come *testing*, e andando a vedere quanti sono i geni che l'algoritmo ha classificato come essenziali quando realmente lo erano



(*True Positives*) e dividendo questa quantità per la somma degli stessi più il numero di geni che sono stati classificati come non essenziali ma che invece lo erano (*False Positives*).
Grazie all'integrazione del software pre-esistente di Weka con la Java Virtual Machine sono riuscito quindi a creare un classificatore Bayesiano all'interno di Java e, dopo aver impostato l'attributo di classe su "Essenziale", ho fornito all'algoritmo bayesiano (che nel programma ho racchiuso in una classe separata chiamata "TestDataSet") iterativamente un diverso data-set, che comprendesse sempre gli stessi dati (100% delle istanze), ma a cui restasse come unico attributo la caratteristica da stimare secondo il parametro PPV.
La procedura di apprendimento usata da parte del classificatore è stata quella della cross-validation a 10 pagine e la Precision viene restituita secondo una funzione implementata da Weka. Risultato di ciò è stato quello che si vede nel grafico: la caratteristica che sovrastima la *Precision* rispetto a tutte le altre è, come anticipato, la ritenzione Filetica; le altre vanno via via a peggiorare nella capacità predittiva, e alcune, come il "*TFBS count*" e "*Upstream size*" ci danno pochissime informazioni.



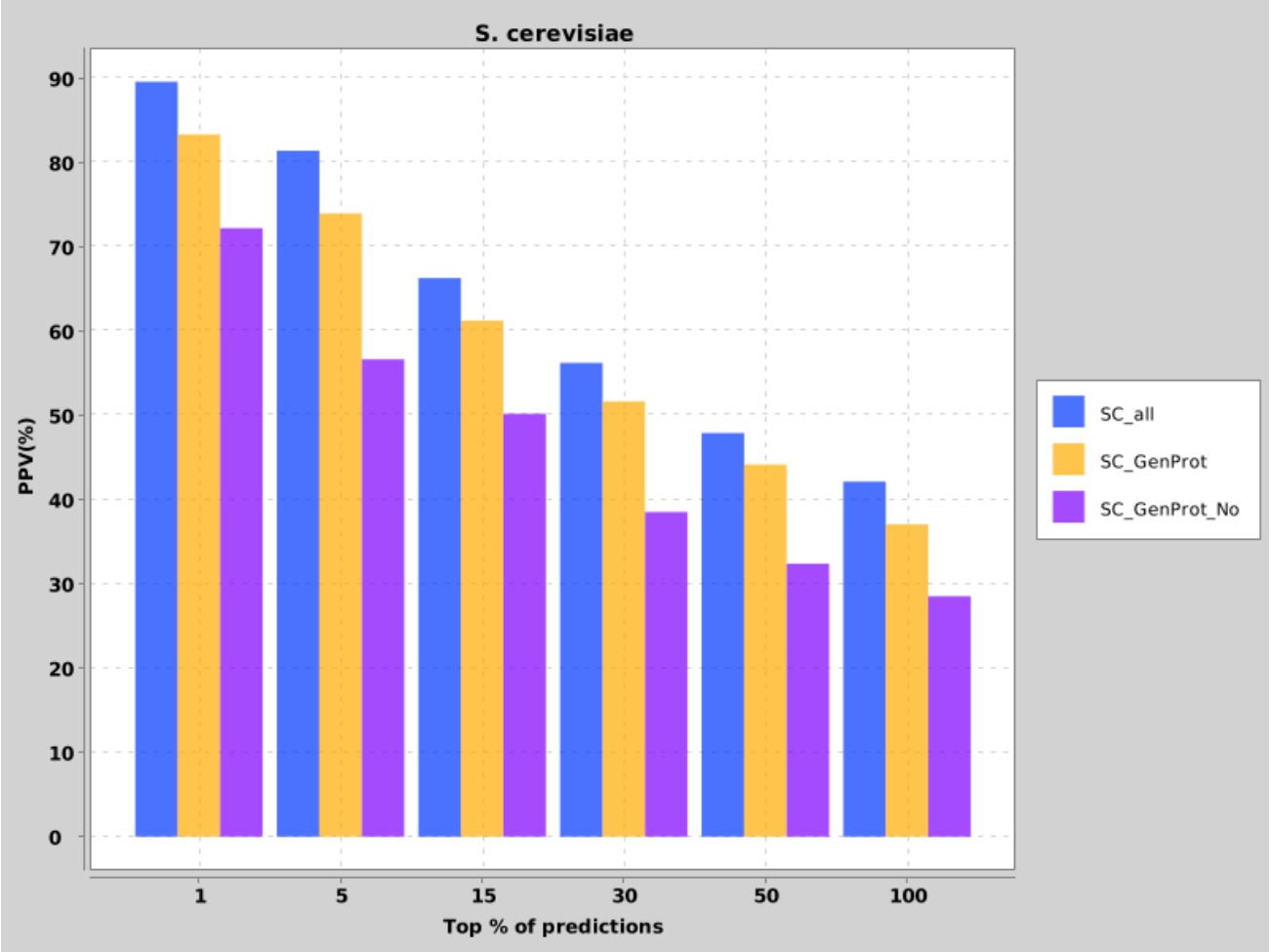
Project4B. Lo stesso procedimento di *ranking* delle caratteristiche l'ho utilizzato anche per il data set di geni di E.Coli, e il risultato è stato questo che si vede nel grafico a sinistra.

Project6A. Il problema successivo è stato quello di stimare diversi set di caratteristiche per *S.Cerevisiae*, sempre secondo il parametro PPV, ma anche di classificarne il risultato in più percentuali di *top* predizioni. I diversi set di caratteristiche corrispondono a:

- DataAll: un set che comprende le prime 23 caratteristiche più predittive
- DataGenProt: un set delle top 11 caratteristiche ottenibili direttamente dalla sequenza genomica
- DataGenProtNo: lo stesso set di GenProt ma con le prime 13 caratteristiche e senza la Ritenzione Filetica

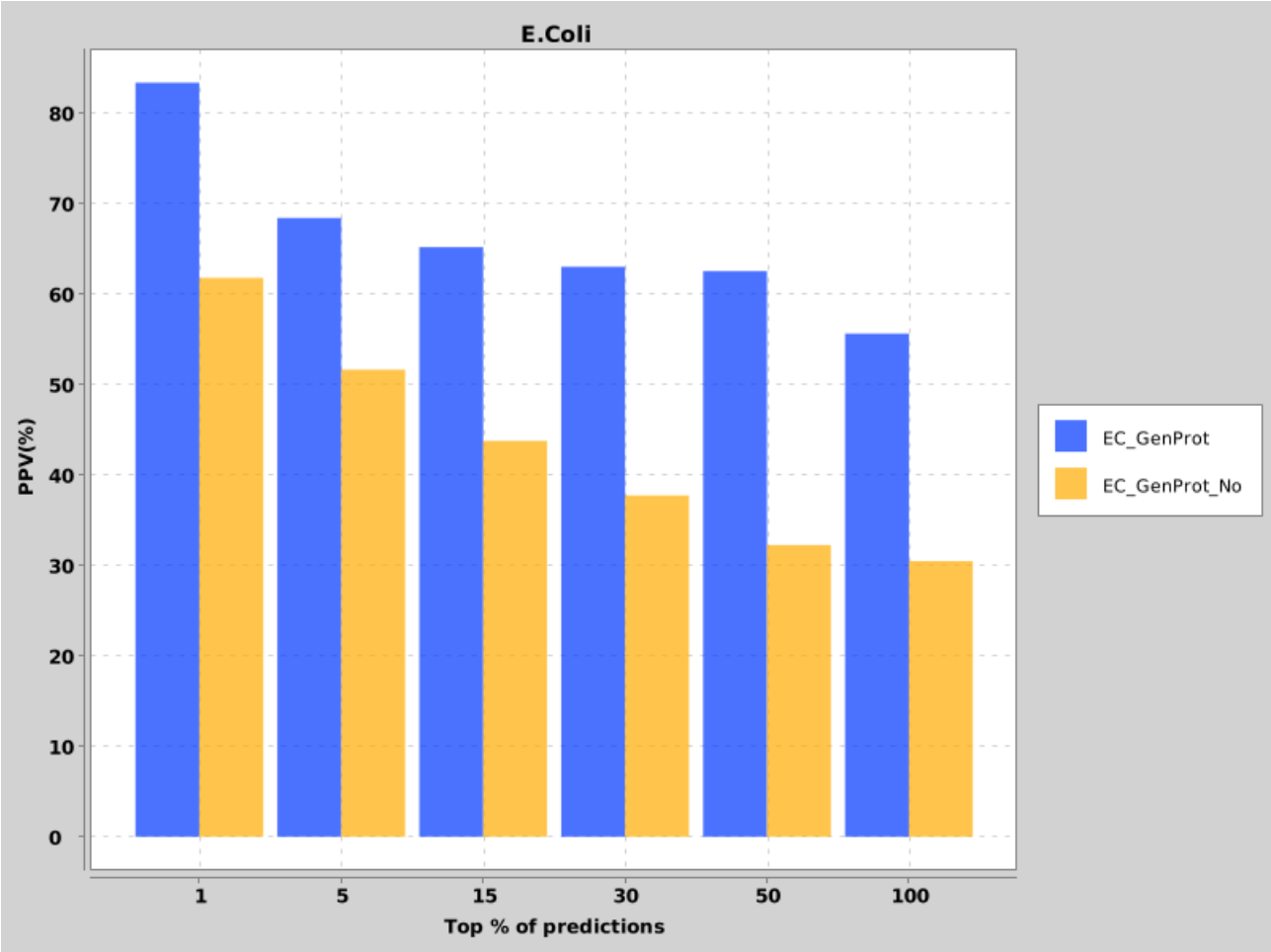
Il procedimento che ho usato è simile a quello dei progetti 4A e 4B ma non del tutto identico: innanzitutto ho fornito al programma 3 diversi data set di geni già costruiti aventi le caratteristiche di cui avevo bisogno, ma in più ho aggiunto degli ulteriori data set, denominati "*Results*", che contengono per ogni frammento di gene l'informazione sulla probabilità di essere essenziale. Ho quindi creato una lista di "Score nominali", cioè una lista di classi che contengono come attributi l'identificatore di ogni gene ed il suo score, che in

questo caso è rappresentato proprio dalla probabilità per lui di essere essenziale. In seguito uso un algoritmo di *sorting* per ordinare la lista secondo lo score, e la "taglio" in base alla *Threshold* considerata (1,5,15,30,50% del numero di istanze), e parallelamente elimino dal data set tutti i geni corrispondenti che sono stati "tagliati fuori" nella lista. Arrivati questo punto il mio data set contiene i geni con più alta probabilità di essere essenziali, cioè quei geni che con facilità mi faranno raggiungere un punteggio più alto nella stima di PPV, dato che sto considerando quelli che con più probabilità potranno essere identificati come "*True Positive*" (Veri Essenziali).
A questo punto applico il classificatore Bayesiano col metodo della cross-validation a 10 pagine al data set così modificato e il risultato di PPV sarà il risultato della Precision di un data set contenente una certa collezione di caratteristiche e un certo tipo di istanze selezionate come migliori ed in una quantità che dipende dalla soglia scelta.
Il risultato è mostrato in figura:



Al solito, il data set che non contiene informazioni sulla caratteristica di Ritenzione Filetica è destinato ad essere meno preciso degli altri per quanto riguarda la predizione.

Project6B. Il procedimento è identico a quello del progetto 6A con l'unica differenza che l'esperimento viene condotto sui data set di E.Coli e non viene considerato il data set "DataAll" perché dà ben poche informazioni aggiuntive rispetto a "DataGenProt".



Nel grafico sovrastante il data set "DataGenProt" contiene le prime 4 top caratteristiche predittive, mentre "DataGenProtNo" contiene le prime 9. Si nota come il dataset DataGenProt, nonostante abbia più informazioni esplicitate dal numero superiore di attributi a sua disposizione, non riesce comunque ad essere più preciso di un dataset con 5 caratteristiche in meno. Questo significa che la caratteristica di ritenzione filetica ha molto più potere predittivo rispetto alle altre sue concorrenti.