Progetto d'esame per il corso di Bioinformatica: Predizione della funzione delle proteine con metodi di Machine Learning

Federico Picetti

Michele Valsesia

Anno accademico 2017/2018

Indice

Capitolo 1

Dati

Le feature di ingresso sono tratte dalla matrice di adiacenza delle proteine di *Drosophila melanogaster*. La matrice esprime una metrica di similarità fra coppie di proteine. Gli algoritmi di apprendimento automatico utilizzano l'i-esima riga (o i-esima colonna) della matrice come vettore di feature per l'i-esimo esempio. Si dispone di 3 distinte matrici di annotazioni, un per ogni ontologia della *GO* (Gene Ontology):

CC Cellular Component, 235 classi

BP Biological Process, 1951 classi

MF Molecular Function, 234 classi

Si tratta di ontologie multiclasse, per cui ogni proteina può appartenere a una o più classi nella stessa ontologia. Le matrici di annotazioni Y riportano le proteine sulle righe e le classi sulle colonne.

$$Y_{i,j} = \begin{cases} 1 & \text{se l'elemento } i\text{-esimo appartiene alla classe } j\text{-esima} \\ 0 & \text{altrimenti} \end{cases}$$

Capitolo 2

Metodi di machine learning

Si è deciso di utilizzare tre diversi metodi di Machine Learning: per Support Vector Machine a AdaBoost si sono utilizzate le librerie scikit-learn per Python. Pegasos è stato implementato in Python in modo da rispettare le API di scikit-learn.

2.1 Support Vector Machine

Si sono provate le SVM della libreria scikit-learn, in particolare l'implementazione SVC[1].

2.2 AdaBoost

2.3 Pegasos

[1] http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

Capitolo 3

Implementazione

Eventualmente parlare di dettagli su come è costruito il codice

Capitolo 4

Risultati

4.1 Metriche adottate

Si è adottata la tecnica di cross-validazione 5-fold. Per ogni classe vengono costruiti 5 classificatori simili e addestrati su 4/5 dei dati. Ogni classificatore viene poi testato sul restante 1/5 dei dati. L'operazione viene eseguita all'interno del modulo `metrics.py`,

4.2 Analisi dei risultati