Metodi di Ensemble Gerarchici per la Predizione Strutturata della Funzione delle Proteine

Studente: Marco ODORE
Matricola: 868906

Relatore: Prof. Giorgio VALENTINI
Correlatore: Dr. Marco NOTARO

Il problema della predizione della funzione delle proteine è uno dei problemi centrali della Biologia Computazionale e presenti diversi problemi aperti sia da un punto visto computazionale sia da un punto di vista biologico.

Uno dei problemi rilevanti in questo ambito è rappresentato dal fatto che ad ogni proteina possono essere associate funzioni molteplici e tali funzioni possono essere rappresentate tramite classe funzionali (termini) a loro volta strutturate secondo un grafo diretto aciclico (DAG). Dal punto di vista dell'apprendimento automatico questo problema può essere modellato con un problema di classificazione multi-label strutturato, in cui le predizioni associate ad ogni proteina sono insiemi di termini della Gene Ontology (GO) strutturate secondo un DAG.

Recentemente sono stati proposti diversi metodi di ensemble gerarchici, progettati proprio per la predizione stutturata dei termini della Gene Ontology (e più in generale per tassonomie organizzate secondo un DAG). Tali metodi si caratterizzano per procedure di learning a due passi, caratterizzate da un primo passo di "flat learning" in cui le classi della GO sono apprese in modo indipendente, e da un secondo passo in cui le predizioni flat sono combinate opportunamente tramite un metodo di ensemble gerarchico.

L'obiettivo primario della tesi consiste nel confrontare metodi di machine learning flat con i metodi di ensemble gerarchici, in modo da valutare se ed in quali condizioni tali metodi possono migliorare significativamente le predizioni flat. Tale confronto è effettuato utilizzando le proteine di un piccolo verme nematode, *Caenorhabditis Elegans*, un organismo modello molto usato per studi di biologia dello sviluppo, il cui genoma è costituito da circa 20000 geni. I risultati mostrano che i metodi gerarchici sono in grado di migliorare sistematicamente le performance di classificazione ottenuti con un ampio spettro di metodi di apprendimento automatico e suggeriscono che tali metodi, sfruttando la loro modularità, potrebbereo essere utilizzati per migliorare le prestazioni di qualsiasi algoritmo di apprendimento flat, a condizione che le predizioni flat su cui si basano gli algoritmi gerarchici non siano quasi totalmente casuali.

Nella tesi viene inoltre proposto un nuovo algoritmo di apprendimento gerarchico basato sulla isotonic regression, chiamato *True Path Rule con Isotonic Regression* (ISO-TPR), basato sull'applicazione dell'algoritmo *Generalized Pool Adjacent Violators* (GPAV) nella fase "top-down" di apprendimento dell' algoritmo di ensemble gerarchico *True Path Rule*. I risultati ottenuti con tale metodo sono risultati essere competitivi con i metodi di ensemble gerarchici allo stato dell'arte per la predizione della funzione delle proteine.

La tesi è suddivisa in diverse sezioni:

- Una parte introduttiva, in cui sono affrontate le criticità e caratteristiche del problema della predizione della funzione delle proteine (capitolo 1).
- Una parte metodologica, in cui si analizzano i metodi ensemble utilizzati allo stato dell'arte, e le nuove proposte (GPAV e ISO-TPR) (capitolo 2).
- Una parte di analisi preliminare del problema specifico di classificazione (Predizione della funzione delle proteine per la specie *C. Elegans*), in cui viene descritto il dataset utilizzato, gli algoritmi di machine learning utilizzati per implementare i base learner dei metodi ensemble, le metriche di valutazione, le stime dei tempi di calcolo, e i metodi per la riduzione della complessità temporale risultante da un problema di classificaione strutturata con migliaia di classi. (Capitolo 3, fino al paragrafo 3.6).
- Una parte sperimentale di classificazione delle classi GO di *C. Elegans*, in cui sono comparati ed applicati i metodi di machine learning flat ed i metodi di ensemble gerarchici su tutto il dataset, valutandone i risultati in relazione a metriche "per-classe" e "per-proteina". (Capitolo 3, dal paragrafo 3.6 fino al 3.9).
- Una parte conclusiva in cui si riassumono i risultati ottenuti, evidenziando in che misura e quando i metodi ensemble sono risultati efficaci e proponendo eventuali lavori futuri. (Capitolo 4).