**Università degli Studi di Salerno**

**Dipartimento di Informatica**

****

**Corso di Laurea in Informatica**

**Classificazione di dati bioinformatici attraverso elementi di Data Science**

|  |  |
| --- | --- |
| **Relatore** | **Candidato** |
| Prof. Michele Risi  Dott.ssa Maria Frasca | Gerardo De Rosa  Matricola: 05121/03762 |

**Anno Accademico 2017/2018**

# Abstract

Negli ultimi decenni lo svilupparsi di tutte le discipline, ed in particolare dell’informatica, ha messo in chiaro che la componente più sfruttata ed importante di qualsiasi studio è rappresentata dai dati; questi sono il motore principale della conoscenza e pertanto risultano fondamentali per qualsivoglia studio si vada ad affrontare.

Per questo motivo è andata formandosi una materia il cui scopo principale è proprio quello di studiare come elaborare e sfruttare i dati, per far sì che questi restituiscono la maggior quantità e qualità possibile di informazioni.

Questa “nuova” tecnologia prende il nome di Big Data e sta prendendo sempre più piede in diversi campi, fornendo un importante ausilio a diverse discipline e permettendone lo sviluppo in maniera esponenziale.

I Big Data utilizzano appunto grandi quantità di dati, la maggior parte delle volte eterogenei, i quali vengono elaborati attraverso diverse tecniche per diversi scopi, tra cui:

* ricavare predizioni sui comportamenti futuri degli stessi, in modo da essere sempre competitivi sul mercato;
* scoprire il legame tra diversi fenomeni per arrivare ad importanti conclusioni in ambito scientifico.

Una delle discipline in cui vengono impiegati in maniera massiccia è la ricerca bioinformatica.

La bioinformatica è una scienza composta da diverse discipline, tra cui l’informatica, la matematica, la biologia e la biochimica; il quale scopo ultimo è quello di analizzare a livello molecolare quesiti principalmente biologici, con strumenti matematici, per cercare di trovare la soluzione al problema analizzato che si avvicini maggiormente a quella ottimale.

Questa scienza in dettaglio si occupa di:

* fornire modelli statistici validi per l'interpretazione dei dati provenienti da esperimenti di biologia molecolare e biochimica per identificare tendenze e leggi numeriche;
* generare nuovi modelli e strumenti matematici per l'analisi di sequenze di DNA, RNA e proteine al fine di creare un corpus di conoscenze relative alla frequenza di sequenze rilevanti, la loro evoluzione ed eventuale funzione;
* organizzare le conoscenze acquisite a livello globale su genoma e proteoma in basi di dati al fine di rendere tali dati accessibili a tutti, e ottimizzare gli algoritmi di ricerca dei dati stessi per migliorarne l'accessibilità.

I dati bioinformatici per definizione, non sono omogenei, per questo motivo e poiché in ogni tipo di elaborazione è sempre presente del “rumore”, non esiste un metodo riconosciuto unanimemente come migliore per lo studio degli stessi e ciò non si riferisce soltanto alla diversità nei campi di studio, ma anche ai campioni di dati, quando questi variano di grandezza e composizione.

Date queste premesse, il punto di partenza di questo elaborato è uno studio effettuato su di un dataset di minori dimensioni ma della stessa tipologia, che tramite un efficace analisi dei dati è riuscito ad ottenere un eccellente differenziazione tra i diversi stati dei campioni.

Gli scopi principali di questa tesi, risultano quindi quelli di:

* analizzare l’algoritmo sviluppato nel precedente elaborato ed il suo comportamento su di un dataset di dimensioni considerevolmente maggiori;
* migliorare l’algoritmo inserendo nuove tecniche di elaborazione, perfezionando la feature selection;
* operare un confronto sui risultati ottenuti, evidenziando le migliorie apportate.

Tutto questo sarà effettuato con l’ausilio di diverse tecniche e metodologie che si avvalgono di strumenti statistici ed informatici.

Il dataset che verrà studiato è composto da liste di campioni formati da geni, prelevati da persone affette da Leucemia; in particolare i campioni esaminati si riferiscono a due categorie di Leucemia: la leucemia mieloide acuta (AML) e leucemia linfoblastica acuta (ALL).

I dati numerici all'interno del dataset sono valori beta, ossia valori che rappresentano le intensità fluorescenti delle sonde in esperimenti a doppio canale di microarray, sulle quali viene effettuata un’operazione di enrichment analysis da gene ontology; questi sono stati scaricati dalla banca dati di GEO[[1]](#footnote-1), il quale è stato sviluppato dall'NCBI come risorsa eterogenea per la sottomissione e il recupero di dati.

Lo studio inizia con un’operazione di ricostruzione del dataset, volte ad eliminare i valori NA non analizzabili, poi si procede con un’analisi dei dati nel loro insieme, volta a identificare le componenti principali del campione dei dati, per poi procedere a ridurre il consistente numero di geni, in modo da semplificare e velocizzare il successivo lavoro di classificazione dei campioni.

Le parti principali che riguardano il pre-processing dei dati comprendono la ricostruzione del dataset, la normalizzazione dei dati e la rimozione degli outlier, che hanno preparato il dataset per l’elaborazione.

La riduzione del numero di variabili è stata volta a classificare i geni in base al loro coefficiente differenziale tra due diversi stati biologici (ALL e AML), ossia in base alla loro significatività statistica relativa alla distinzione tra i due stati; per ridurre poi ulteriormente l’insieme delle variabili, tramite l’uso di una libreria di R che permette l’identificazione dei geni che maggiormente concorrono alla differenziazione dei tipi leucemici.

A seguito della riduzione del numero di variabili, sono stati implementati dei modelli di classificazione con l’utilizzo delle reti neurali, delle macchine a vettori di supporto e delle reti residuali, in modo da analizzare le differenze tra i modelli computazionali in relazione a questo insieme di dati.

1. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi [↑](#footnote-ref-1)