Università degli studi di Milano-Bicocca

High Dimensional Data Analysis

Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes

Authors:
Lorenzo Camaione - 850380- l.camaione@campus.unimib.it
Martino Lorenzo Tenconi - MTS4128991- m.tenconi3@campus.unimib.it
Lorenzo Famiglini - MTS4128991- l.famiglini@campus.unimib.it

April 22, 2020

Abstract

The ABSTRACT is not a part of the body of the report itself. Rather, the abstract is a brief summary of the report contents that is often separately circulated so potential readers can decide whether to read the report. The abstract should very concisely summarize the whole report: why it was written, what was discovered or developed, and what is claimed to be the significance of the effort. The abstract does not include figures or tables, and only the most significant numerical values or results should be given.

1 Introduction

La dieta influenza significativamente la salute umana ed in parte anche il microbioma dell'intestino. In questo studio, da un campione di 97 persone, sono stati estratti dati sulla loro dieta e la composizione del DNA delle loro feci. I batteri presenti nelle feci possono essere clusterizzati in enterotipi che differiscono tra loro perlopiù per i livelli di Bacteroides e Prevotella. Inoltre, questi enterotipi possono essere associati con diete assunte per un lungo periodo. In particolare alti livelli di Bacteroides possono essere associati a diete ricche di proteine e grasso animale mentre alti livelli di Prevotella con carboidrati. Oltre ad analizzare l'associazione dieta-enterotipo si sono anche studiati dei modelli che spiegano il livello di BMI in funzione di nutrienti, microbioma e fattori demografici.

2 Datasets

I dataset a disposizione per le analisi sono forniti dal portale QIITA (study ID 1011) e sono rappresentati da una tabella OTU, una tavola tassonomica e un questionario FFQ relativo ai nutrienti assunti dai soggetti nella dieta a lungo termine. La tabella OTU, di dimensioni 3393 x 100, contiene il numero di sequenze che sono state osservate per ognuno dei 3393 batteri presenti nel microbioma di ciascuno dei 100 pazienti. Ognuno dei batteri può essere presente o assente nel microbioma di ogni paziente e la combinazione di questi valori determina il microbioma vero e proprio di ciascun paziente. La tavola tassonomica permette di collegare gli identificativi dei batteri presenti nella tabella OTU alle loro tassonomie; in questa tavola sono indicati per ognuno dei 3393 batteri i sette livelli tassonomici regno, phylum, classe, ordine, famiglia, genere e specie. Infine il questionario dei nutrienti contiene, per 98 soggetti, le quantità di ciascuno dei nutrienti assunti in una dieta di lungo periodo. In aggiunta a questi tre dataset vi è una quarta tabella contenente gli id dei 100 soggetti in studio e una serie di variabili, come ad esempio il genere, l'altezza, il peso, l'età e il Body Mass Index (BMI) a loro associate. Le informazioni estratte da questo ultimo dataset, combinate con le variabili estratte nella fase di preprocessing, saranno utilizzate per scopi di analisi predittiva del BMI.

2.1 Preprocessing

Si è effettuata una prima pulizia della tavola tassonomica rimuovendo caratteri speciali dai nomi delle tassonomie e si sono analizzate quindi la tavola e le osservazioni per ciascuna colonna indicante il livello tassonomico. Poiché lo studio [referenza paper] è improntato all'analisi dei batteri a livello di genere, ci si concentra sull'analisi delle variabili indicanti i nomi delle tassonomie fino al sesto livello, ignorando la specie dei batteri osservati. La tavola tassonomica presenta un totale di 1526 batteri con genere non identificato e dunque ai fini della analisi si considerano per questi batteri i soli livelli precedenti presenti; ad esempio per un batterio che ha classificazione fino al quarto livello ("ordine") si considerano unicamente le prime quattro tassonomie eliminando quindi i valori mancanti per i successivi livelli. A questo punto vengono raggruppati tra loro i batteri nella tabella OTU che hanno

la stessa classificazione fino all'ultimo livello presente; si ottengono quindi cinque dataset, ognuno per ogni livello tassonomico da phylum a genere, contenenti le possibili aggregazioni osservate per il relativo livello. Per ognuna di queste cinque tabelle sono calcolate le abbondanze relative dei batteri, ovvero la presenza in termini percentuali di ciascun batterio in ognuno dei 100 individui considerati. Aggregando queste cinque tabelle si ottengono 198 possibili livelli di aggregazione dei batteri osservati. Sono state considerate, in seguito, solamente le tassonomie presenti in almeno il 10% dei soggetti che avessero un valore di abbondanza relativa maggiore o uguale allo 0.2% in almeno uno dei soggetti. In questo modo sono state ottenute 76 tassonomie. Inoltre, per scopi successivi nelle analisi, si decide di creare una tabella OTU contenente i batteri aggregati al solo livello di genere o superiore ove mancante (famiglia, genere, classe, ...), ottenendo così 119 tassonomie differenti. Si mostrano le abbondanze nei 97 pazienti dei generi di batteri osservati:

figura

Per quanto riguarda il dataset dei nutrienti, di dimensione 98 x 215, sono state selezionate soltanto le 97 righe con identificativo del soggetto nello studio presente anche nella tabella OTU citata in precedenza. Per valutare quali nutrienti fossero maggiormente significativi a livello di impatto sulla composizione del intestinale, è stata considerato l'intero insieme dei 3393 diversi batteri non aggregati per livello tassonomico come indicatore di variabilità della composizione del microbioma stesso. Sulle abbondanze della tabella OTU originaria sono state dunque calcolate le distanze tra i microbiomi intestinali dei 97 soggetti con la metrica Unifrac non pesata, individuata come metodo maggiormente discriminante per le differenze di composizione dei microbiomi rispetto alla stessa metrica Unifrac pesata, che tiene quindi in considerazione la quantità di ciascun batterio nel microbioma, piuttosto che la semplice presenza/assenza come nel caso della metrica non pesata. Il calcolo della distanza utilizzando la Unifrac non pesata avviene tramite la formula:

$$\frac{sum\ of\ unshared\ branch\ lengths}{sum\ of\ all\ tree\ branch\ lengths} = fraction\ of\ total\ unshared\ branch\ lengths$$

Una volta calcolata la matrice delle distanze non pesate Unifrac sui 97 microbiomi intestinali, è stato possibile effettuare la PERMANOVA su ciascuno dei nutrienti standardizzati della tabella. A differenza della ANOVA, che si basa sull'ipotesi di normalità, la PERMANOVA traccia test di significatività confrontando il risultato del test F reale con quello ottenuto dalle permutazioni casuali degli oggetti tra i gruppi senza fare alcuna assunzione sulla distribuzione dei dati; inoltre, mentre ANOVA verifica la somiglianza delle medie tra i gruppi basandosi sui dati, la PERMANOVA utilizza in input una matrice di distanze. Ognuno dei nutrienti è stato valutato significativo se il valore di False Discovery Rate FDR (percentuale di volte in cui si rifiuta l'ipotesi nulla quando invece è vera) è minore al 25%. Come descritto nel paper di riferimento, una questa soglia relativamente alta è stata utilizzata per non escludere dall'analisi nutrienti associati a batteri con valori di abbondanza relativa non particolarmente alti. I batteri e i nutrienti selezionati sono stati utilizzati per costruire due dataset con lo scopo di effettuare l'analisi predittiva del BMI. Sono state estratte le prime 6 componenti principali dei nutrienti selezionati (varianza catturata pari a x%) e le prime 5 per quanto riguarda le tassonomie rilevanti, con un x% di varianza catturata. Le differenze tra i 2 dataset sono esclusivamente dovute alle tassonomie dei batteri considerati: mentre nel primo dataset sono state considerate tutte le 119 tassonomie a livello di genere o superiore se mancante, nel secondo dataset sono state inserite solamente le tassonomie a livello di genere, senza considerare potenziali livelli superiori. Le componenti principali presenti nei due dataset sono state estratte da queste due diverse aggregazioni. I dataset utilizzati per l'analisi predittiva sono stati suddivisi in training e test: la parte dedicata all'addestramento dei modelli è risultata essere il x%, lasciando il x% dei dati per la validazione dei modelli.

3 The Methodological Approach

This is the central and most important section of the report. Its objective must be to show, with linearity and clarity, the steps that have led to the definition of a decision model. The description of the working hypotheses, confirmed or denied, can be found in this section together with the description of the subsequent refining processes of the models. Comparisons between different models (e.g. heuristics vs. optimal models) in terms of quality of solutions, their explainability and execution times are welcome.

Do not attempt to describe all the code in the system, and do not include large pieces of code in this section, use pseudo-code where necessary. Complete source code should be provided separately (in Appendixes, as separated material or as a link to an on-line repo). Instead pick out and describe just the pieces of code which, for example:

- are especially critical to the operation of the system;
- you feel might be of particular interest to the reader for some reason;
- illustrate a non-standard or innovative way of implementing an algorithm, data structure, etc..

You should also mention any unforeseen problems you encountered when implementing the system and how and to what extent you overcame them. Common problems are: difficulties involving existing software.

4 Results and Evaluation

The Results section is dedicated to presenting the actual results (i.e. measured and calculated quantities), not to discussing their meaning or interpretation. The results should be summarized using appropriate Tables and Figures (graphs or schematics). Every Figure and Table should have a legend that describes concisely what is contained or shown. Figure legends go below the figure, table legends above the table. Throughout the report, but especially in this section, pay attention to reporting numbers with an appropriate number of significant figures.

5 Discussion

The discussion section aims at interpreting the results in light of the project's objectives. The most important goal of this section is to interpret the results so that the reader is informed of the insight or answers that the results provide. This section should also present an evaluation of the particular approach taken by the group. For example: Based on the results, how could the experimental procedure be improved? What additional, future work may be warranted? What recommendations can be drawn?

6 Conclusions

Conclusions should summarize the central points made in the Discussion section, reinforcing for the reader the value and implications of the work. If the results were not definitive, specific future work that may be needed can be (briefly) described. The conclusions should never contain "surprises". Therefore, any conclusions should be based on observations and data already discussed. It is considered extremely bad form to introduce new data in the conclusions.

References

The references section should contain complete citations following standard form. The references should be numbered and listed in the order they were cited in the body of the report. In the text of the report, a particular reference can be cited by using a numerical number in brackets as [**?**] that corresponds to its number in the reference list. LaTeXprovides several styles to format the references