Alternative Methods in Longitudinal Data Analysis and Prediction Modeling

Advanced Topics in Data Mining and Machine Learning | Jad Habouch | February 21, 2018

1. READINGS

Two-part Mixed-effects Model For Analyzing Longitudinal Microbiome Compositional Data, Bioinformatics, Chen, 2016, Oxford Academic:
https://academic.oup.com/bioinformatics/article/32/17/2611/2450750

Longitudinal Prediction of the Infant Gut Microbiome with Dynamic Bayesian Networks, McGeachie, 2015, Scientific Reports: https://www.nature.com/articles/srep20359

2. ABSTRACT

Longitudinal data (sometimes referred to as panel data) analysis focuses on understanding how a subject of interest evolves at different points in time. At a broader scale and in practice, we collect longitudinal data by measuring a variable of interest recurrently over time for a group of subjects. This enables longitudinal data to combine the properties of both cross-sectional data and time-series data. The standard approaches for analyzing this type of data have included modeling the expected value of the response variable as either a linear or nonlinear function of a set of explanatory variables (sas.com, 2018). When applying these methods and analysis techniques to model more complex and highly dynamic subjects of interests, these traditional models have many limitations. In this presentation, we will explore alternative methods working with longitudinal data modeling, with the goal of prediction improvements.

Both the Chen et al, and McGeachie papers use microbiome and microbiota as their data domains. The collection of all the microorganisms including; bacteria, archaea, protists, fungi and viruses, that live in association with the human body is termed the “Human Microbiota”. These microorganisms inhabit human tissues and bio-fluids, including the skin, mammary glands, placenta, seminal fluid, uterus, ovarian follicles, lung, saliva, oral mucosa, conjunctiva, biliary and gastrointestinal tracts. Microbiome is the collective genomes of resident microorganisms.

Chen et al paper demonstrates “a two-part zero-inflated Beta regression model with random effects (ZIBR) for testing the association between microbial abundance and clinical covariates for longitudinal microbiome data”. The second method from the McGeachie et. al paper constructs a Dynamic Bayesian Networks to model the interaction of many microbial species in the gut microbiome. The goal of both methods is creating better predictive modeling capabilities.

3. **SPOTLIGHT QUESTION:** Are machine learning methods and techniques agnostic to the data domain and types? Are they better suited for certain types of data and questions? Does their complexity in implementation and modeling always justify their benefits?