DOCTORAL THESIS

Learning Phenotypes from Electronic Health Records Using Robust Temporal Tensor Factorization

YIN, Kejing

Date of Award:
2021

Abstract

With the widespread adoption of electronic health records (EHR), a large volume of EHR data has been accumulated, providing researchers and clinicians with valuable opportunities to accelerate clinical research and to improve the quality of care by advanced analysis of the EHR data. One approach to transforming the raw EHR to actionable insights is computational phenotyping — the process of discovering meaningful combinations of clinical items, e.g. diagnosis and medications, from the raw EHR data for characterizing health conditions with minimum human supervision. Many data-driven approaches have been proposed to tackle the problem, among which non-negative tensor factorization (NTF) has been shown effective for high-throughput discovery of phenotypes from structural EHR data.

Although great efforts have been made, several open challenges limit the robustness of existing NTF-based computational phenotyping models. (1) The correspondence information between different modalities (e.g., between diagnosis and medication) is often not recorded in EHR data, and existing models rely on unrealistic assumptions to construct input tensors for phenotyping which introduces inevitable errors. (2) EHR data are often recorded over time, presenting serious temporal irregularity: patients have different lengths of stay and the time gap between clinical visits can vary significantly. Existing models are limited in considering the temporal irregularity and temporal dependency, which limits their generalizability and robustness. (3) Heavy missingness is unavoidable in the raw EHR data due to recording mistakes or operational reasons. Existing models mostly do not take the missing data into account and assume that the data are fully observed, which can

greatly compromise their robustness.

In this thesis research study, we propose a series of robust tensor factorization models to address these challenges. First, we propose a hidden interaction tensor factorization (HITF) model to discover the inter-modal correspondence jointly with the learning of latent phenotypes. It is further extended to the multi-modal setting by the collective hidden interaction tensor factorization (cHITF) framework. Second, we propose a collective non-negative tensor factorization (CNTF) model to extract phenotypes from temporally irregular EHR data and separate phenotypes that appear at different stages of the disease progression. Third, we propose a temporally dependent PARAFAC2 factorization (TedPar) model to further capture the temporal dependency between phenotypes by capturing the transitions between them over time. Forth, we propose a logistic PARAFAC2 factorization (LogPar) model to jointly complete the one-class missing data in the binary irregular tensor and learn phenotypes from it. Finally, we propose context-aware time series imputation (CATSI) to capture the overall health condition of patients and use it to guide the imputation of clinical time series.

We empirically validate the proposed models using a number of real-world, large-scale, and de-identified EHR datasets. The empirical evaluation results show that the proposed models are significantly more robust than the existing ones. Evaluated by the clinician, HITF and cHITF discovers more clinically meaningful inter-modal correspondence, CNTF learns phenotypes that better separate early and later stages of disease progression, TedPar captures meaningful phenotype transition patterns, and LogPar also derives clinically meaningful phenotypes. Quantitatively, LogPar and CATSI show significant improvement than baselines in tensor completion and time series imputation, respectively. Besides, HITF, cHITF, CNTF, and LogPar all significantly outperform baseline models in terms of downstream prediction tasks.

Keywords: Computational phenotyping, non-negative tensor factorization, binary tensor completion, PARAFAC2 factorization, time series imputation.

Table of Contents

List of Tables

List of Figures