CHIL-M
CHallenges in analyzing Incomplete Longitudinal Medical data

TU/e supervisor: Rianne M. Schouten, TU/e
Additional supervision by: dr. Wouter Duivesteijn, TU/e
Additional supervision by: prof.dr. Mykola Pechenizkiy, TU/e
Additional supervision by: dr. Jolanda J. Luime, Erasmus MC

Introduction

In this project, we work together with the Dutch south-west Early Psoriatic Arthritic Registry (DEPAR), which is a collaboration of 15 medical centers in the Netherlands that aim to investigate which patient characteristics, measurements and scales are useful in determining an appropriate treatment for patients diagnosed with Psoriatic Arthritis (PsA), see `https://ciceroreumatologie.nl/depar`.

The registry collects all sorts of information from patients diagnosed with PsA: baseline characteristics, anamnesis, physical examination, results from blood tests, etcetera. Data collection occurs during every hospital visit (5 times in the first year, 2 times in the second year, and every subsequent year). The resulting dataset is heterogeneous, longitudinal, contains lots of missing data and is challenging to analyze.

Possible directions

The project is divided into two sub-projects (and has place for two students):

1. **Handling missing values:**

   The DEPAR dataset has been analyzed to investigate clinical outcomes such as disease activity, psoriasis burden and health-related quality of life [6, 2, 3]. Most of these studies use information from patients where data is fully observed.

   However, DEPAR also contains many incomplete records. In this project, we aim to find and apply valid and suitable missing data methods for analyzing this partially observed data. Consequently, we will investigate missing data patterns, learn about missing data solutions, propose an approach and evaluate the effect of that approach (for an introduction to Missing Data (MD) methodology, see [5]).

   This project has a strong applied component: we follow up on medical research in the field of PsA, will have to understand the data and research questions from a clinical point of view and need to propose missing data solutions that are not only valid but also practical and interpretable.

2. **Discovering subgroups with exceptional longitudinal patterns:**

   Current DEPAR studies analyze the *entire* sample of PsA patients to investigate characteristics, measurements and scales for patients with PsA. For instance, [6] use a mixed-effects model to analyze disease activity measures over the course of 1 year, and [3] use Principal Component Analysis (PCA) to find latent traits that could help rheumatologists to distinguish PsA patients with a high psoriasis burden from those with a low burden.

   In this project, we use the framework of Exceptional Model Mining (EMM) [1], a local pattern mining technique, to search for *group-specific* effects. EMM strives to find interesting subgroups in a dataset.

We find subgroups interesting if they satisfy two conditions. On the one hand, they must be interpretable: we must be able to define subgroups as a conjunction of a few conditions on columns of the original dataset. On the other hand, they must be exceptional: some kind of interaction (the model class) between several target columns of the dataset must have exceptional parameters. The simplest example of such an interaction is exceptional correlation between two numeric targets: on a dataset describing the housing market, EMM could then find certain segments of the housing market where the correlation between lot size and sales price disappear.

We aim to develop a model class for longitudinal data. Some work in this direction has been done (some work for mixed-effects models yet has to be released, and some work on repeated cross-sectional data can be found in [4]), but more directions can be explored.

Requirements

In this project, you will represent TU/e in a collaboration with rheumatologists and researchers from 15 hospitals in the Netherlands. It will be necessary to interact with clinicians in order to make your research clinically meaningful. Furthermore, the data is highly secured and requires careful treatment. You will have to sign an NDA and become part of the DEPAR research team as a guest researcher. Therefore, we require a professional attitude.

Furthermore, solving MD problems and developing a new EMM model class both require some affinity with statistics. Willingness to learn more about statistics is fine as well.

When done well, we expect the EMM project to lead to a publication in a data mining conference or journal. The MD project may lead to a publication in an applied journal that is closely related to the medical domain.

References

[1] Wouter Duivesteijn, Ad J Feelders, and Arno Knobbe. Exceptional model mining. *Data Mining and Knowledge Discovery*, 30(1):47–98, 2016.

[2] Fazira R Kasiem, Marc R Kok, Jolanda J Luime, Ilja Tchetverikov, Kim Wervers, Lindy-Anne Korswagen, Nastasja HAM Denissen, Yvonne PM Goekoop-Ruiterman, Maikel van Oosterhout, Faouzia Fodili, et al. The burden of psoriasis in patients with early psoriatic arthritis. *Rheumatology*, 61(4):1570–1578, 2021.

[3] Fazira R Kasiem, A Pasma, JJ Luime, I Tchetverikov, K. Wervers, LA Korswagen, N Denissen, Goekoop-Ruiterman YPM, M van Oosterhout, Fodili F., JMW Hazes, MBA Van Doorn, MR Kok, and M Vis. A practical guide for the assessment of psoriasis burden in patients with psoriatic arthritis. *The Journal of Rheumatology*, 2022.

[4] Rianne M Schouten, Wouter Duivesteijn, and Mykola Pechenizkiy. Exceptional Model Mining for Repeated Cross-Sectional Data (EMM-RCS). In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*, pages 585–593. SIAM, 2022.

[5] S. Van Buuren. *Flexible imputation of missing data*. CRC press, 2018.

[6] Kim Wervers, Jolanda J Luime, Ilja Tchetverikov, Andreas H Gerards, Marc R Kok, Cathelijne WY Appels, Wiebo L van der Graaff, Johannes HLM van Groenendael, Lindy-Anne Korswagen, Josien J Veris-van Dieren, et al. Comparison of disease activity measures in early psoriatic arthritis in usual care. *Rheumatology*, 58(12):2251–2259, 2019.