Diabetes profiling

1. Project idea

Diabetes affects approximately 415 million people globally, but its treatment still relies on the outdated knowledge that a patient is type I or II, meaning at the extreme end of the diagnostic spectrum. It is now acknowledged that diabetes is a more versatile disease and most of patients have a profile which lies somewhere between type I and II. Our project aims to provide a new method for diabetic profiling by analysing synthetic diabetic patient data, which potentially enables more personalized care and thus fewer complications and reduced health care costs.

2. The story behind the project

Diabetes is traditionally divided into etiology based types I and II. However, type I and II only represent the extreme ends of the diabetic spectrum with the vast majority of patients having symptoms and markers from both subgroups. Currently, diagnosing patients into these specific subgroups is based on subtype definitions and therefore the entire process is inherently vague. In addition, most patients transition between different diabetic profiles as their disease proceeds.

We wanted to develop a profiling method relying on a patient data lake with more than 20k patients, 1.5M prescriptions and 13M measured lab values. Our aim was to enable more customized patient care by discovering (more or less) distinct profiles for patients with diabetes. The profiles would provide patients and professionals information about risk factors associated with the respective profile and their relevant drug prescriptions.

Our profiling method is based on the assumption that patients with high glucose measurements (B-Gluk, fP-Gluk, P-Gluk and B-HbA1c) and poor glucose balance are more susceptible to complications resulting from hyperglycemia. These complications can be

divided into microvascular (nephro-, neuro- and retinopathic complications) and macrovascular (e.g. coronary artery disease and stroke). These analyses would be most effective in time series form. Our profiling method also takes into consideration the effect of smoking and heavy drinking.

However, for legal reasons the time series data relating to measurements and medication prescriptions given to us in this challenge was entirely randomized and we decided not to try to present it in a meaningful way. Our data analysis pipeline consists of accessing data lake with Apache Spark, cleaning and shaping the desired data into a patient indexed matrix. We then train a neural network of linear dense layers using Tensorflow backend on few of the most important lab values from the matrix and age at death, to get a neural predictor for life expectancy. This predictive model is a) importable to many front-ends b) re-trainable and c) expandable upon new data points. We build a small widget for demoing purposes.

In Finland, diabetes care makes up for 15% of the total annual healthcare costs, majority of which relate to treatment of hyperglycemic complications. Approximately two thirds of complications resulting from hyperglycemia are preventable. With appropriate data for measurement time, our project could be further developed to predict future progression of the disease based on what their diabetic profile is now and what it has been like in the past.

We also explored ways of using the dataset for uses that are not directly linked to diabetes. We wanted to know which medications are often prescribed to same patients. This would allow to study e.g. adverse effects of drug interactions. The most relevant results were visualized with an interactive heat map.

Anton Mattsson, Esa Turkulainen, Lauri Pykälä, Milja Leinonen and Santeri Mentu