Clustering and prediction of electronic health record data from mental health patients in a Finnish healthcare environment

21.5.2019

Electronic health records (EHR) contain a wealth of information of interest to both the patient and the service provider but are historically not designed for easy computational analysis. To further complicate the matter the data set is often unstructured, contains both numerical and categorical variables and is very large in size. Thus far studies on EHR data has been limited to small-scale studies, and few large-scale studies on different machine learning methods have been conducted on such data.

In his master’s thesis, Oskar Niemenoja, who majors in Complex systems, explores three different clustering methods and two predictive models on a large set of EHR data of Finnish occupational healthcare patients from January 2012 to February 2019, studying people that have used mental health services during that period. He further introduces a concept of treatment pathways in a medical context as treatment sessions related to one single initial diagnosis. Within this scope he explores the viability of Gower’s dissimilarity matrix with hierarchical clustering, factor analysis with mixed data (FAMD) with k-means clustering, and model-based method to label different mental health related treatment pathways in the data. Based on these labels he explores the viability of random forest and support vector machine classifiers to predict various outcome variables on the data from the initial visit. These outcome variables contain for example the length of the treatment path or the probability that the person will be needing sick leave during his treatment, which are interesting aspects to both the patient and healthcare provider for effective planning of treatment.

In his work Niemenoja demonstrated how these clustering and predictive models work on health record data and validated the results statistically and with expert evaluation. As a result he finds that different clustering methods produce very different results in terms of the size and number of different diagnoses contained in a treatment pathway. The error rates for these models ranged from a few percent to more than 30% according to two medical experts. He further explains how these different models can be used in different contexts to create diverse insight into the treatment paths of mental health patients. The predictive models are proven to be an adequate tool to predict the treatment pathway and sick leave lengths. Additionally, these methods are excellent at finding out unusually large values for these measures making them useful in identifying patients at risk early on in the treatment pathway.

Overall Niemenoja demonstrates that the selected methods are a suitable tool for analyzing large-scale EHR data and provides results for both clustering and prediction models as well as assessment on the quality of such clusters and predictions. Overall the results are promising, working as an initial study into further studies into further structuring and predicting large-scale medical data.

More information:
Oskar Niemenoja
contact@oskarniemenoja.fi