Håvard Slettvold

Clustering of Patient Trajectories

Master project, fall 2014

Patient Trajectories project
Department of Computer and Information Science
Faculty of Information Technology, Mathematics and Electrical Engineering

Abstract

Health care is an enormously complex area, and it requires patients, staff and data to combine efforts across several institutions to carry out diagnosis and treatment. While many institutions today use electronic journals to keep track of single patients, not many systems are being used to collect data about patients with similar properties and use that data as part of the clinical process. The EHR is a potential source for great information that could benefit decision making and perhaps even help patients understand how their treatment is progressing compared to other cases.

Extrapolation of coherent data from the raw format of these journals usually starts by attempting to identify similarities and differences between patients to produce patterns that identify a grouping. Such processing is called *clustering*. PAsTAs has a few projects that aim to improve clinical care, and as a part of that project some form of clustering will be the basis for future research into chronic conditions. The data used for this thesis is extracted from anonymised EHJ of patients in the greater Trondheim area and the condition they have in common is diabetes.

Using previously published articles about clustering of patient data will be essential to the process. An analytical approach will also require inquiry with clinical care staff so proper boundaries can be defined for what entails a relevant aspect of a patient trajectory. Finally understanding the field of clustering will prove essential to creating patterns to carry out the calculation of clusters to be used in a visualisation solution.

Preface

This research is the work of my master thesis written autumn 2014 at the Department of Computer and Information Science (IDI) at the Norwegian University of Science and Technology (NTNU). The project is part of the Patient Trajectories project: Clustering of Patient Trajectories, lead by Rune Sætre at IDI.

Data used as the basis for this thesis is a result of a unique approval from REC[1] to gather patient journals from hospitals, general practitioners and care institutions in Trondheim, Melhus, Midtre Gauldal and Malvik. Staff Engineer Håkon Dale Wågbø has been responsible for extracting and processing the data previous to the usage in this project.

The supervisors for this thesis have been Associate Professor Rune Sætre and Associate Professor Øysten Nytrø. Another actor in this project is Aslak Steinsbekk, Professor in Behavioural sciences in medicine and Health service research at the Department of Public Health and General Practice at NTNU.

Håvard Slettvold
Trondheim, December 20, 2014

[1] Regional Committees for Medical and Health Research Ethics

Contents

List of Figures

List of Tables

Chapter 1

Glossary

NTNU Norwegian University of Science and technology.

NST Norwegian Centre for Integrated Care and Telemedicine.

REC Regional Committees for Medical and Health Research Ethics.

IDI Department of Computer and Information Science at NTNU.

PAsTAs PAtient's TrAjectories; A cooperative project between NST and NTNU.

EHJ Electronic Healthcare Journal.

EHR Electronic Healthcare Registry. This registry contains instances of EHJ.

CDS Clinical Decision Support.

Chapter 2

Introduction

PAsTAs is a cooperative project between the Norwegian Centre for Integrated Care and Telemedicine (NST) and the Norwegian University of Science and Technology (NTNU), where;

The project aims to collect EPJ-data from general practitioners and hospitals, including care and care services, and display these together in an event stream. This registry will be the foundation for out research on trajectories on both individual and aggregated levels.[1]

This thesis will focus on Clustering within that scope. While this thesis will focus on dividing lager groups of patients into groups based on similarity of a given set of conditions, the overall goal of the patient trajectories project is to allow better care for patients by improving patient trajectories in general for primary care.

2.1 Background

Large amounts of data can be found in the EHR. The great challenge of using this data is to determine which patient records contains similarities. Acquiring data from hospitals, general practitioners and other public care services and analysing data across their registries will allow us aggregate a substantial base of data to properly understand how treatment unfolds over time.

Electronic Healthcare Registries (EHR) are quite common today, many countries use it to store patient data, which is the basis for Clinical Decision Support (CDS) and aggregation of data for other purposes. Introducing EHR into clinical care to enable CDS is a good motivating factor. In Osheroff et al. [2007],

[1] Quote from http://www.telemed.no/pastas-pasientforloep.5219575-247951.html (Visited 2014-12-15)

it's shown that it holds great value for clinicians, patients and other health care stakeholders to have properly filtered information at appropriate times in the clinical process.

2.2 Motivation

It's also shown that computer-based, clinician-directed CDS systems found that over 90% of the systems significantly improved clinical care in randomized controlled trials, provided that the CDS was made available automatically as a part of clinician workflow, offered an actionable recommendation, and was delivered at the time and location of clinical decision making [Kawamoto et al. [2005]].

Improving the quality of the data presented to clinicians at these stages can be improved from the ground up, by providing hard clusters containing as revevant data as possible. Studies on scalable CDS shows that there is a lack of standards capable of defining data representation, knowledge representation and mapping [Kawamoto et al. [2010]]. These problems translates down to the clustering stages, a problem which is easily visible in the raw data available for this thesis.

2.3 Goals and Research Questions

The focus of this project will be to analyse the patient trajectories of cases pertaining to chronic diseases. At the end of this project I should be able to visualise patient trajectories by combining data from several institution registries in a greater area.

Goal *Find a reasonable way to make clusters of patients based on relevant parameters from their condition, care or outcomes.*

Diabetes will be the chronic disease that the clustering will focus on. Grouping such trajectories together will require development of patters capable of producing clusters. This will include creating a solution for scalable visualisation of aggregated data based on different filters or scales. Two such filters could be:

- Relation of group. Meaning you can adjust the trajectories based on severity of condition, age or other distinguishable parameters from patients.
- Size of institution. Meaning adjustments will change whether to include a whole hospital as one entity, split it into different departments or wards, or even specific practices.

Research question 1 *How to define a pattern which creates relevant clusters?*

Choosing how to group patients and treatment institutions together requires understanding of how the healthcare sector works, as well as understanding how treatment works.

Research question 2 *How to visualise data expressing trending or abnormal trajectories?*

Demonstration of the different groupings will require a visual representation. Creating a logically sound way to display how the different groupings affect behaviour of trajectories will be important.

2.4 Research Method

The research method will be analytical in nature, and using literature searches to uncover previous projects and how they have tackled the problem of clustering.

Understanding which parameters are most vital to a diabetes trajectory is key when attempting to create the patterns used for clustering. Consulting with clinical care staff regarding regular behaviour of a diabetes sickness trajectory is important when assigning values for each data point. Solving the first research question will involve talking to clinical staff about disease patterns and professors about practices in clustering.

Aggregation of data based on defined filters requires a well structured data model. Analysing patterns of patient trajectories to uncover trending or abnormal trajectories within these sets of data will allow a visualisation to express them in the best way possible.

2.5 Thesis Structure

The thesis consists of five main chapters:

- Chapter 2: introduces the project and the research goals
- Chapter 3: Information about the fields relevant to the research
- Chapter 4: Shows the scientific models used in the research
- Chapter 5: Presents the results of the research
- Chapter 6: Evaluates the results and a plan for the future research

After the main chapters comes a list of bibliography.

Chapter 3

Background Theory and Motivation

In this chapter the theory for the technologies and relevant fields of the thesis will be discussed. The contents are just theory meant to introduce terminology, and will not contain discussions about method or results.

3.1 Background Theory

To understand the method and terminology of the thesis it is important to understand the fields which the thesis is built upon. In the coming sections these will be explained in sufficient detail to understand the methods used in the process of clustering.

3.1.1 Information Retrieval

As a base for clustering, it is important to understand how a program will interact with a data source. In Information Retrieval (IR) we call one source of data a "document". Documents can in reality be any kind of file, not just written text, but also images, audio or video files. When and IR-system calculates how relevant each document is to another it uses some for of aggregated data from these documents and adds them to an index, which is used when a user comes up with queries to find relevant information.

EHJ could easily be used as a data source in this manner. The data we receive from REC doesn't come in the form of journals, but as raw data with each line representing an entry into some institutions system indicating that a patient has

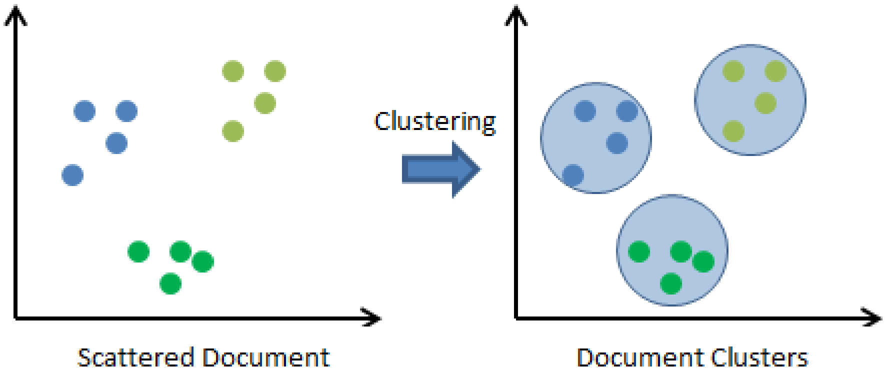

Figure 3.1: Example of how clustering can be represented in a vector graph (source: codeproject.com)

visited. Processing this data into an improvised journal and then using that "document" as the basis for clustering is probably the best approach.

3.1.2 Clustering

"Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters)" [Jain et al. [1999]].

Clustering is not a new concept, the term originated in the 1970s and considerable studies have been made into this area. Some studies have also been made into how different patient trajectories can be grouped based on similarity of symptoms, diagnosis, treatment and mortality [Jensen et al. [2014]].

Intuitively, patterns within a valid cluster are more similar to each other than they are to a pattern belonging to a different cluster. An example is depicted in figure 3.1. In this example each document is represented as a two dimensional vector. The similarities between each document is then decided based on how near the vectors are to each other. This is a common way of organising similarities in IR, and the vector can have as many dimensions as is necessary to produce differences.

When looking at patterns of data, one cluster should intuitively contain patterns that are different to those of another cluster. Defining algorithms that work out which patterns seem similar is a very complex task. Defining clusters are usually done in the following steps [Jain and Dubes [1988]]:

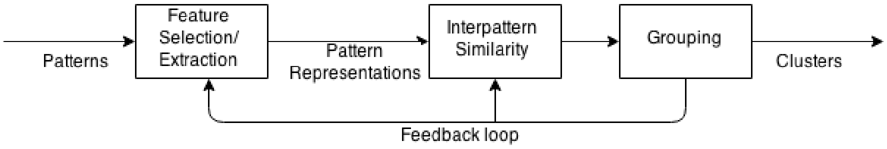

Figure 3.2: Showing steps 1-3 in the clustering process (source Jain et al. [1999])

1. Pattern representation (optionally including feature extraction and selection)
2. Definition of a pattern proximity measure appropriate to the data domain
3. Clustering or grouping
4. Data abstraction (if needed)
5. Assessment of output (if needed)

(1) Pattern representation patterns are shown in the left side of figure 3.1, and the desired clusters can be seen on the right side. When processing each EHJ, it could be possible to represent the data in such a way, by assigning weights to factors in the journal.

(2) Pattern proximity is determined by examining two patterns. Distance measures can be as simple as measuring euclidean distance between two patterns, which best reflects dissimilarities. More advanced measures can be used to calculate the conceptual similarity between two patterns.

(3) The grouping step attempts to separate the produced data points into clusters. In Jensen et al. [2014], a study of the Danish EHR, the Jaccard Index was used for this step. Jaccard Index compares similarity and diversity of two sets, which could be attributes for each document.

(4) Data abstraction is the process of extracting a simple and compact representation of a data set. Either as data to be used in further processing or as human readable data that is easy to comprehend.

(5) Evaluating output from clustering algorithms is important in order to refine the pattern representation for future processing of data. It's important to note that clustering algorithms will find clusters in any kind of data, even if there are none. The hard part is to define a good enough pattern so that you actually get a spread in the data which produces a significant dissimilarity between each document and still produce good clusters.

Chapter 4

Method

This chapter will explain the scientific model for this thesis. The methods were chosen based on a literature review and previous experience.

4.1 Structured Literature Review

Literature for this project was in large part of the curriculum for TDT38 - Clinical decision- and process support. This theory module encompasses some of the medical area this thesis needs to cover. Almost any kind of system that uses CDS needs some form of clustering to evaluate similarity and relevance. Some suggestions have also been given by the supervisors.

For literature on clustering Google Scholar[1] was used. Searches as simple as "clustering" revealed articles which explains the theory and progress of clustering very well.

4.2 Initial parameters

The visualisation for the clustering is supposed to filter or display some different values based on a few given parameters. The adjustments for the visualisation should allow a user to specify the granularity of clustering, where the most broad scope would be to check whether a patient has any record of hospital appointments and the most specific would be each individual service.

Examples of the scopes in order from broad to more specific can be seen in table 4.1. Splitting the municipal services into daily and hourly services is done to specify certain categories that we know to appear in EPJ entries in the data

[1] http://scholar.google.no/

1	Hospital - admissions and polyclinic
2	Municipal services - split into daily and hourly services
3	General Practitioners
4	Emergency ward
5	Other services appearing in the data

Table 4.1: The general layout for the institution hierarchy to be used for filtering

that is available, which is a sizeable amount. By grouping these together in the hierarchy it will improve readability and seem more intuitive.

4.3 Development model

Considering the development of the application as a whole, including both clustering and visualisation, will most likely depend quite a bit on feedback before yielding desired results, an adaptive development model is desirable.

Many adaptive models are already commonly being used in the software industry, often labeled “agile” . While these methodologies certainly have proven to efficiently handle feedback and changing design in stride, providing a solid framework for groups of developers to work together on larger projects and improving reliability it is probably wasted on a project with only one developer.

Instead some of the principles from Martin [2003] will be used to create a productive cycle that allows for regular input. As shown in figure 4.1 this is how each task will be carried out, although not every step from this model may be present in every task.

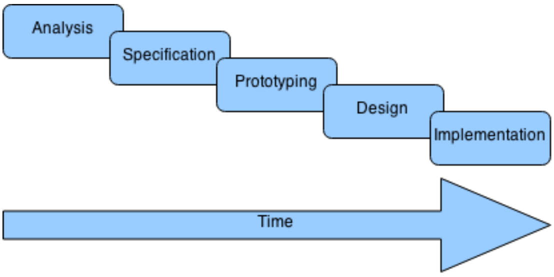

Figure 4.1: Illustration showing the development model

Agile development strategies also build on another important principle; to divide any task into subtasks, which should all be reduced to the smallest size that still fulfils some requirement of a greater task. This division of each task means that it should be easier to see the entire backlog[2] of tasks and compare it to what is already being done and what is completed, in order to establish a good impression of overall progress.

[2] Planned tasks that are not yet started

Chapter 5

Results

In this chapter the preliminary results of the chosen models for this thesis will be discussed.

5.1 Structured Literature Review

This section will outline some thoughts about the articles found by the SLR process.

5.1.1 Clustering

To answer the primary research goal of this thesis, an understanding of the creation of patterns to be used in clustering is essential. There are multitudes of articles and books on this subject, most of which are closely tied to fields such as artificial intelligence (AI) and information retrieval (IR).

My experience in the AI field is quite limited, but it is my understanding that some elements of general clustering problems can be avoided if the domain is well specified, and the desired outcomes are not wildly dissimilar. In order to create the patterns needed for this project, I believe a decent understanding of IR and how similarity is calculated, primarily in text documents, can help create a sort of clustering process that can be quickly updated and filtered on the fly.

Clustering is explained well in the two articles Jain and Dubes [1988] and Jain et al. [1999]. According to the historical summary of clustering, it mostly requires a refining process that is guided by parameters which over each iteration makes sure that each cluster is defined as “hard” as possible, meaning that it attempts to trim fuzzy patterns away from the cluster and improve the overall similarity.

5.1.2 Usages in healthcare

A few of the articles found in the curriculum for TDT38 are relevant to what this project is trying to accomplish. In Jensen et al. [2014], the clustering process is not grouping patient trajectories, but rather trying to group disease trajectories. The process might be similar to what I need, but I believe the Jaccard Index used in this research is not sufficient to separate individual patient trajectories when they all pertain to the same condition.

Quite a few articles in the curriculum also mentions temporal data abstraction as an important method to gather point and interval information and arrive at unified qualitative descriptions of parameters [Miksch et al. [1996]]. The process of temporal abstraction is explained in greater detail in an article by Stacey and McGregor [2007], where the temporal abstraction parameters are tested against live patient data to provide detection of patterns within multiple patient data streams.

Chapter 6

Evaluation

In this chapter there will be some summarising of the progress so far, and what is planned for the coming part of the project in the spring.

6.1 Evaluation

I joined the PAsTAs project in late September, meaning that work on other parts of the project already had already been going on for a few months. Considering the time disadvantage I will have to conduct more research in the spring that should have happened during the fall.

6.2 Discussion

The report for the fall does not present many tangible results to answer the research goals, but rather builds up the foundation for the work that will happen in the spring. Contact with clinical staff in the diabetes field should have been made during the fall, but this will certainly be one of the first points on the agenda for the final work in the spring.

6.3 Future Work

Starting in the spring semester will begin with contacting the endocrinology department at St. Olavs hospital to investigate the common disease trajectories for diabetes and map significant data points based on that feedback.

I also expect that some more research needs to be made into the field of clustering, as I have little experience with the process. Understanding how to define proper pattern is key to producing eligible clusters for a future project.

Bibliography

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for clustering data.* Prentice-Hall, Inc.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323.

Jensen, A. B., Moseley, P. L., Oprea, T. I., Ellesøe, S. G., Eriksson, R., Schmock, H., Jensen, P. B., Jensen, L. J., and Brunak, S. (2014). Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nature communications*, 5.

Kawamoto, K., Del Fiol, G., Lobach, D. F., and Jenders, R. A. (2010). Standards for scalable clinical decision support: need, current and emerging standards, gaps, and proposal for progress. *The open medical informatics journal*, 4:235.

Kawamoto, K., Houlihan, C. A., Balas, E. A., and Lobach, D. F. (2005). Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *Bmj*, 330(7494):765.

Martin, R. C. (2003). *Agile software development: principles, patterns, and practices.* Prentice Hall PTR.

Miksch, S., Horn, W., Popow, C., and Paky, F. (1996). Utilizing temporal data abstraction for data validation and therapy planning for artificially ventilated newborn infants. *Artificial intelligence in medicine*, 8(6):543–576.

Osheroff, J. A., Teich, J. M., Middleton, B., Steen, E. B., Wright, A., and Detmer, D. E. (2007). A roadmap for national action on clinical decision support. *Journal of the American medical informatics association*, 14(2):141–145.

Stacey, M. and McGregor, C. (2007). Temporal abstraction in intelligent clinical data analysis: A survey. *Artificial intelligence in medicine*, 39(1):1–24.