Patient Treatment Pathways Clustering

Ahmed Najjar[1], **Christian Gagné**[1], **and Daniel Reinharz**[2]
[1]Laboratoire de vision et systèmes numériques, Dép. de génie électrique et de génie informatique
[2]Laboratoire de simulation du dépistage, Dép. de médecine sociale et préventive
Université Laval, Québec (Québec), Canada
ahmed.najjar.1@ulaval.ca, christian.gagne@gel.ulaval.ca
daniel.reinharz@fmed.ulaval.ca

Abstract

Clustering electronic medical records allows discovery of information on health care practises. Entries in such medical records are usually made of a succession of diagnostics or therapeutic steps. The corresponding processes are complex and heterogeneous since they depend on medical knowledge integrating clinical guidelines, physicians individual experience, and patient data and conditions. To analyze such data, we are first proposing to cluster medical visits, consultations, and hospital stays into homogeneous groups, and then to construct higher-level patient trajectories over these different groups. These patient trajectories are then also clustered to distill typical pathways, enabling interpretation of clusters by experts. This approach is evaluated on a real-world administrative database of elderly people in Québec suffering from health failures.

Our project stems from three main observations. First, medical treatment processes consist in successions of diagnostics or therapeutic steps linked to a patient. Being able to analyze and to assess these processes have its importance in the healthcare domain. Second, administrative healthcare databases contain observational data, collected for purposes other than data analysis. They have seldom been used for analyzing medical treatment processes, although they are a rich source of information on health services and associated processes provided to patients. And third, process clustering has received increased attention over the years since it allows the clustering of the execution traces contained in an event log generated by many latent processes. Being able to extract families of general models representing the health services provided from the administrative databases would allow the comparison of real-life medicine with the official guidelines, providing, for example, a capacity to identify good practices that stem from physician know-how.

In recent years, some proposals have been made for the clustering of medical processes. For instance, Rebuge et al. [1] proposed a methodology based on first-order Markov chain to cluster processes composed of the care events occurring in the emergency department of a hospital. Whereas, Huang et al. [2, 3] have applied latent Dirichlet allocation (LDA) to discover latent patterns as a probabilistic combination of clinical activities. They assume that a patient clinical pathway is represented by a mixture of treatment patterns. They apply LDA to two specific care flow logs concerning intracranial hemorrhage and cerebral infarction extracted from a hospital information system. The model gives the clinical activity density estimation for each pattern, from which the probabilistic association between an activity and a pattern can be obtained. These works on process clustering of health data rely on processes made of relatively simple and well-defined events. In contrast, the real-life relational administrative databases we are using for our project contain numerous tables and links, which should be first transformed into process logs, as a succession of medical services.

These services are characterized by their categories and associated variables. To reduce the complexity of the processes logs, we are proposing to cluster these complex objects, so as to relabel them into categories and clusters. The clustering of the services depends on variables types. For data characterized only by categorical variables, hierarchical clustering over one variable at a time is done.

Service	Physician specialty	Diagnostic	Intervention
s73 (59.3%)	General pract. (98.0%)	Left heart fail. (42.9%)	None (55.1%)
s78 (23.7%)	General pract. (81.3%)	Left heart fail. (43.8%)	No chir. colonoscopy (25.0%)
v92 (23.7%)	General pract. (81.1%)	Heart fail. (14.2%)	–

Table 1: Description of cluster 16, composed of 59 instances. Values in parenthesis represents the support of the element in the cluster.

Representative instance	Pathway	Number of neighbours
Most central instance	{s73}	21
Second most central instance	{s78,v172}	11

Table 2: Two most central instances representing cluster 16.

If datasets are described by mixed types, with numerical, categorical, and multivalued categorical variables, we apply the mixed values k-prototypes algorithm we previously proposed in [4].

From these clustering results, process logs are represented as a succession of complex object labels. Hidden Markov Models (HMMs) are then used to cluster them, using the method proposed by Knab et al. [5]. With this method, one HMM is associated to each cluster, with a clustering scheme organized in two steps. The first step consists in assigning each log sequence of a patient to the HMM with the highest emission probability. At the second step, the HMMs are updated according to the log sequences assigned to them. Based on sequences and data assignment conducted at the first step, HMM parameters are recomputed using the Baum-Welch algorithm. These two steps are repeated until convergence or resource exhaustion.

To analyze the results, an algorithm is proposed to find representatives for each cluster, relying on the notion of neighbourhood and coverage. For that, we compute the number of neighbours for each process in each cluster using the cosine distance. Processes are then sorted in descending order of number of neighbours and added successively to the representative set until the coverage percentage exceeds a given threshold.

As experimental validation, we are evaluating the proposed approach on the clustering treatment processes of elderly patients over 65 years old who live in the province of Québec (Canada) and are suffering from heart diseases, through the access to subsets of the administrative health care databases from the RAMQ (universal health insurer for Québec residents) and the MSSS (Ministry of Health). We have preprocessed these databases by gathering various medical services according to three categories: visits, consultations, and hospital stays.

For our experiments, 10,000 individuals were selected, such that we have extracted 38,102 hospital stays, 156,149 consultations and 749,964 visits conducted between January 1, 2000 and December 31, 2009. Visits and consultations are clustered based on age, sex, and diagnosis chapter variables, using hierarchical cluster over one variable at a time. For hospital stay datasets, objects are characterized by numerical, categorical and multivalued categorical variables. We thus first split objects by age and sex, and after we apply our mixed values k-prototypes algorithm [4]. This gave us 10,000 patient treatment processes represented as a succession of medical services labels. The likelihood of HMM clustering results is used to determine the proper number of clusters, the sweeping varying between 10 to 20 clusters, with $k = 18$ clusters selected given it maximizes the measure.

It is interesting to note that, although pathways are complex, our approach has allowed us to discover different clusters within the health processes. Therefore, it become possible to cluster a huge amount of processes into homogeneous clusters and to see latent patterns and their characteristics. For example, Table 1 provides a description of cluster 16 obtained with our approach, characterized by some specific hospital stay for cardiac disease (s73, s78) and visit for cardiac disease (v92). This cluster contains 61.02 % pathways formed by only one hospital stay. Moreover, from 560 possible different complex object labels, the pathways of this cluster were composed of only 44 different labels. Table 2 represents the two most central instances of this cluster according to their number of neighbours. Results obtained demonstrate that the proposed approach allows a differentiation between pathway clusters.

References

[1] Álvaro Rebuge and Diogo R Ferreira. Business process analysis in healthcare environments: A methodology based on process mining. *Information Systems*, 37(2):99–116, 2012.

[2] Zhengxing Huang, Xudong Lu, and Huilong Duan. Latent treatment pattern discovery for clinical processes. *Journal of medical systems*, 37(2):1–10, 2013.

[3] Zhengxing Huang, Wei Dong, Lei Ji, Chenxi Gan, Xudong Lu, and Huilong Duan. Discovery of clinical pathway patterns from event logs using probabilistic topic models. *Journal of biomedical informatics*, 47:39–57, 2014.

[4] Ahmed Najjar, Christian Gagné, and Daniel Reinharz. A novel mixed values k-prototypes algorithm with application to health care databases mining. In *IEEE Symposium Series on Computational Intelligence (IEEE-SSCI)*, pages 159–166. IEEE, 2014.

[5] Bernhard Knab, Alexander Schliep, Barthel Steckemetz, and Bernd Wichern. Model-based clustering with hidden Markov models and its application to financial time-series data. In *Between Data Science and Applied Data Analysis*, pages 561–569. Springer, 2003.