Proceedings of the 13th IFAC Symposium on
Information Control Problems in Manufacturing
Moscow, Russia, June 3-5, 2009

Clustering of patient trajectories with an auto-stopped bisecting K-Medoids algorithm *

Hongying Fei, Nadine Meskens, Claire-Hélène Moreau

Louvain School of Management and Catholic University of Mons, 151 chaussée de Binche, B-7000 Mons, Belgium (e-mail: {fei, meskens, moreauc}@ fucam.ac.be)

Abstract: Nowadays, more and more hospitals seek to employ well-developed process management and simulation tools in the healthcare organizations to develop the overall patient pathways. Since it is extremely important to define suitable patient groups for constructing process or simulation models, we proposed a sequence mining method, an auto-stopped Bisecting K-Medoids clustering algorithm, to classify patients into groups with homogeneous trajectories within two stages. At the first stage, patients are classified according to the complexity of the care process. Afterwards, groups obtained at the first stage are further classified with the similarity of the trajectories. The proposed approach was executed with a real data set from a medium-size hospital. According to the experimental results, this method can be used to classify patients into manageable groups, where the most frequent trajectories might be extracted to validate a process modelling technique. In addition, data extracted from those groups could be used to feed our simulation models.

Keywords: Patient trajectory, classification, bisecting K-Medoids, sequence mining

1. INTRODUCTION

Clinical pathways, which are also known as care pathways, critical pathways, integrated care pathways or care maps (Cardoen and Demeulemeester (2007)), are normally used to systematically plan and follow up the treatment process of a patient, i.e. a patient trajectory. Since the 1980s, the Integrated Care Pathways (ICPs) have been introduced into healthcare organizations as structured case-notes to facilitate clinical effectiveness and quality improvement, especially for the multidisciplinary groups. Nowadays more and more countries have started to get involved in implementing ICPs in their hospitals (Hindle and Yazbeck (2005); De Bleser et al. (2006)). Furthermore, we find that in the past two decades, not only a variety of process modelling techniques have been introduced from the industrial system to the healthcare environment to conduct qualitative analysis (Framinan et al. (2005); Ren et al. (2007)), such as the coordination among involved services and the efficiency of communications, but also many discrete-event simulation models have been constructed to make the quantitative evaluation (Ramis et al. (2001); Cardoen and Demeulemeester (2007); Niu et al. (2007)), such as the assignment of suitable professionists and the performance of different management strategies for involved services. In one word, more and more hospitals seek to develop the overall patient trajectories.

The study of existing methodologies and techniques for process modelling and discrete-simulation shows that the most important work is to define groups of patients with homogeneous trajectories. Since the patient trajectories are known to be highly diverse and hard to be standardized (Hall et al. (2006)), it is necessary to develop suitable clustering tools for this kind of patient classification problem. According to the literature, some achievements about patient classification problems were announced but only two of them were found classifying patients by their treatment process (Isken and Rajagopalan (2002); Maruster et al. (2005)). Furthermore, since the latter two studies classify the patients with some extracted characteristics instead of the original information of their trajectories, we doubt whether those obtained groups really consist of patient with homogeneous trajectories or not. Therefore, in this study, we aim to classify patients with original information of their trajectories. In addition, since the patients passing through the operating theatre are the most important part, we will just focus on the development of hospital performance related to this surgical suite. In this study, we aim at classifying patients, who pass through the operating theatre, into groups with homogeneous trajectories.

In general, trajectories of patients passing through the operating theatre consist of three phases: The pre-operative phase, per-operative phase and the post-operative phase (Fei and Meskens (2007)). According to hospital practices, most of the patient's activities take place at the pre-operative phase, and it is obvious that the effectiveness of the pre-operative treatment is important to the arrangement of the operating theatre. Therefore, the sequence involved in our study consists of only pre-operative activities, i.e. the activities from the first visit at the outpatient clinic until the surgery in the operating room.

---

* This study is funded by the "National Fund for Scientific Researcher (FNRS)" and the "Inter-University Attraction Poles Programme" of the Belgian State, Science Policy Unit.

The rest of this paper is organized as follows. In section 2, we explain at first the selection of clustering methods and then introduce the general structure of the method proposed for the considered patient classification problem together with the stop criterion. Afterwards, the measures of similarity for both stages, are introduced in section 3. In section 4, after introducing how the experimental data were collected, we report experimental results and conduct the related analysis. This paper is ended up with some concluding remarks and perspectives.

2. GENERAL PROCEDURE OF THE METHOD PROPOSED FOR THE CONSIDERED PATIENT CLASSIFICATION PROBLEM

2.1 Selection of sequence clustering methods

Since the trajectory of a patient is in fact a sequence of activities that can be symbolized in alphanumeric string, the sequence clustering methods can be implemented to obtain patient groups with homogeneous trajectories. According to the literature, there are in general three categories of sequence clustering strategies (Xu and Wunsch (2005)):

(1) Sequence similarity: this scheme is based on the measurement of the distance (similarity/dissimilarity) between each pair of sequences. Therefore, proximity-based clustering algorithms, either hierarchical or partitional, can be used to group sequences.
(2) Indirect sequence: an indirect strategy is implemented here. In this strategy, all the sequences are then mapped into the transformed feature space, where classical vector-space-based clustering algorithms can be used to form clusters.
(3) Statistical sequence clustering: this approach aims to construct statistical models to describe the dynamics of each group of sequences.

In this study, we are interested in the first two methodologies because they are mainly used to deal with sequential data composed of alphanumeric characters, while the third one is normally applied to numerical or categorical sequences (Xu and Wunsch (2005)). The further analysis shows that the first scheme is quite computationally laborious and hence not suitable for the large-scale sequence databases. The second approach greatly reduces the computational complexity by means of feature extraction; meanwhile it inevitably causes the loss of some information in the original sequences. Considering that in general several thousands of sequences are involved in one medium-size hospital per year, we try to compromise between the effectiveness and efficiency of the algorithm. In consequence, our patient classification problem is solved in two stages. At the first stage, the indirect sequence scheme is implemented so that a set of features are extracted from the involved trajectories to express the complexity of the care process and a quantitative vector, whose elements correspond to the extracted features, is constructed for each patient. Based on those vectors, a clustering method can be employed to classify patients according to the distances between every two quantitative vectors. At the second stage the sequence similarity scheme is used and each group, obtained from the first stage, is further clustered according to the similarity of trajectories. Since the clustering methods used at both stage have similar structures except for the measurement of similarity (the first is used for the quantitative vectors while the second is employed for the alphanumerical sequences), in this study, we aim not only to find proper measures of similarity, but also to propose a general structure of clustering method that is suitable for both stages. Before the details of the former, we would like to introduce at first the latter in this section.

According to the literature, it is found that clustering algorithms, based on K-means method (MacQueen (1967)), are widely implemented to solve practical problems even for the large data sets (Tan et al. (2005); Xu and Wunsch (2005)), therefore we attempt to use this kind of algorithm to solve the considered patient classification problem. Considering still some disadvantages with respect to the development of K-Means algorithms, an auto-stopped bisecting K-Medoids is proposed in this study to overcome two of its drawbacks that are listed as follows.

- No efficient and universal method has been developed for identifying the initial partitions and the number of clusters. Instead of an a priori number K, the bisecting K-Means algorithm (Boley (1998))together with a stop criterion is implemented in our clustering algorithm to directly determine the final clustering results;
- The definition of “means” limits the application only to numerical variables. Since the K-Medoids algorithm has the similar procedure as the K-Means algorithm except that the medoid, i.e. the most representative data point, is used instead of the quantitative means, which do not need any computation and always exist (Kaufman and Rousseeuw (1990)), it is employed in this study.

In the following sections, we would like to firstly introduce the general procedure of the proposed clustering algorithm and the implemented stop criterion.

2.2 General procedure of the proposed clustering algorithm

The general procedure of the proposed bisecting K-Medoids algorithm is as follows.

Step 1: Initialize the list of clusters to contain one universal cluster containing all objects;

Step 2: Select a cluster with the largest intra-distance, i.e. sum of the distance between each object from that cluster and the cluster medoid, and remove it from the current list of clusters. It should be noted that in this algorithm, the medoid of one cluster represents the object that has the lowest cumulative distance to the other objects in this cluster.

Step 3: Bisect the selected cluster using the basic K-Medoids algorithm within a certain trial (in this paper, the basic K-Medoids algorithm is executed 200 times per iteration):

Step 3_1: Select $K (K = 2)$ points as the initial medoids;

Step 3_2: Assign all objects to the closest medoid, i.e. the medoid with the smallest distance.