ARTICLE IN PRESS

Journal of Biomedical Informatics xxx (2013) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Discovery of clinical pathway patterns from event logs using probabilistic topic models

Zhengxing Huang [a,1], Wei Dong [b,1], Lei Ji [c], Chenxi Gan [a], Xudong Lu [a], Huilong Duan [a,*]

[a] College of Biomedical Engineering and Instrument Science, Zhejiang University, Zhou Yiqing Building 510, Zheda Road 38#, Hangzhou, Zhejiang 310008, China
[b] Department of Cardiology, Chinese PLA General Hospital, China
[c] IT Department, Chinese PLA General Hospital, China

ARTICLE INFO

Article history:
Received 14 February 2013
Accepted 7 September 2013
Available online xxxx

Keywords:
Clinical pathway analysis
Topic models
Latent Dirichlet Allocation
Pattern discovery
Clinical event log

ABSTRACT

Discovery of clinical pathway (CP) patterns has experienced increased attention over the years due to its importance for revealing the structure, semantics and dynamics of CPs, and to its usefulness for providing clinicians with explicit knowledge which can be directly used to guide treatment activities of individual patients. Generally, discovery of CP patterns is a challenging task as treatment behaviors in CPs often have a large variability depending on factors such as time, location and patient individual. Based on the assumption that CP patterns can be derived from clinical event logs which usually record various treatment activities in CP executions, this study proposes a novel approach to CP pattern discovery by modeling CPs using mixtures of an extension to the Latent Dirichlet Allocation family that jointly models various treatment activities and their occurring time stamps in CPs. Clinical case studies are performed to evaluate the proposed approach via real-world data sets recording typical treatment behaviors in patient careflow. The obtained results demonstrate the suitability of the proposed approach for CP pattern discovery, and indicate the promise in research efforts related to CP analysis and optimization.

1. Introduction

In order to increase the quality of care services in an unfavorable economic scenario and under the financial pressure by governments, health-care organizations have to introduce clearly defined clinical pathways (CPs) for patients, and these pathways must be improved continuously [1–5]. A CP is a defined set of therapy and treatment activities that represent the steps required to achieve a specific treatment objective in patient careflow. It has been proven that CPs can break functional boundaries and offer an explicit process-oriented view of health-care where the efficient collaboration and coordination of physicians become the crucial issue [6,2,7–13]. Since actual treatment activities are extremely complex, with numerous variations across various stages in CPs, they often bear no relation to the ideal as envisaged by the designers of CPs. To this end, CP analysis is nonetheless vital for health-care management due to its usefulness for capturing the actionable knowledge to administrate, automate, and schedule the best practice for individual patients in the executions of CPs [14–19].

Traditionally, health-care practitioners analyze CPs by looking at aggregated data seen from an external perspective, e.g., length of stay (LOS), charge, bed utilization, medical service levels, and so on [20]. A new and promising way of acquiring insights into CPs is to analyze clinical event logs using process mining techniques [21–24], which have been widely studied in the domain of business process management, which attempt to extract non-trivial and useful information from event logs [24,16]. Regarding clinical settings, many hospital information systems usually record various kinds of treatment events in patient careflow, and each event refers to a specific case (i.e., a patient trace following a specific CP), is related to a particular clinical activity type (i.e., a well-defined step in a specific CP), has an associated occurring time stamp, and other properties. Clinical event logs conceal an untapped reservoir of knowledge about the way of specific therapy and treatment activities being performed on particular patients in their careflow. It is, therefore, possible to mine clinical event logs, extract non-trivial knowledge from these logs, and exploit these for further analysis.

In this study, we focus on discovering CP patterns from clinical event logs. We define CP patterns to be temporal regularities of CPs. A CP often involves several underlying patterns of treatment activities from admission to discharge, possibly over different time scales and for varying time intervals. The underlying patterns form the backbone of CPs and should be conserved, and the absence of presence of such patterns may indicate the cause of an anomaly or a malfunction (e.g., medical error) [25]. Therefore, we argue that discovery of CP patterns offers potentially rich information about patient treatment intent and behaviors, and can provide a basis for further CP analysis.

* Corresponding author. Fax: +86 571 87951960.
E-mail address: duanhl@zju.edu.cn (H. Duan).
[1] These authors contributed equally to this work.

1532-0464/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.09.003

However, discovery of CP patterns is in general a challenging task as the diversity and complexity of treatment behaviors in CPs are far higher than that of common business processes [16]. Although it is possible to employ existing process mining techniques in mining CP patterns from clinical event logs, many of them often generate spaghetti-like patterns which are difficult to be comprehended by clinical experts and either not amenable or lack of assistance to efforts of CP analysis and improvement. In our previous work [19], we developed a new process mining algorithm to derive a concise summary from a clinical event log. While the constructed summary is capable of describing the entire structure of a CP, it is challenged by describing underlying treatment patterns and handling the different levels of complexity of these patterns. Note that what makes the discovery of CP patterns complex is that they are typically composed of various and heterogeneous treatment activities and the composition of activities has a large variability depending on factors such as time, location and patient individual.

To this end, we propose to leverage the power of probabilistic topic models (1) to extract latent CP patterns from event logs, and (2) to enable the recognition of patient traces as a composition of such patterns. In our previous work [18], we adapted a probabilistic topic model, Latent Dirichlet Allocation (LDA) to build a discovery model of latent CP patterns. LDA is an unsupervised probabilistic clustering technique used to discover latent topics from bags of words in text by finding co-occurrences of words in documents. Here, treatment activities are extracted and are mapped to words. These are then collated for a specific patient trace and become analogous to a document. The latent topics discovered by LDA in this way are interpreted as CP patterns. The probability distribution derived from LDA surmises the essential features of CP patterns, and CPs can be accurately described by combining different classes of distributions [18].

While interpretable and meaningful CP patterns can be discovered, our previous work has a major shortcoming that it does not take the occurring time stamps of treatment activities into account. Note that a CP pattern captures not only the low-dimensional structure of treatment activities, but also when these activities are performed in patient careflow. Standard LDA does not distinguish differences in occurring time of treatment activities in CPs. While the occurring time stamp is often critical to capturing the meaning of treatment behaviors in CP analysis, neglecting it may result in spurious associations and misleading inference on capturing the meaning of underlying CP patterns from clinical event logs. Although it is possible to take pairs of "treatment activity–time stamp" as words in topic models, it will not be able to discriminate the generation of treatment activities and that of occurrences.

In this work, we exploit to extend our previous work using LDA in [18] to propose an extension of LDA, i.e., clinical pathway model (CPM), show that richer CP patterns can be automatically discovered from clinical event logs. The proposed model finds latent CP patterns that are influenced by both treatment activities and their occurring time stamps in an unsupervised manner, thus disclosing temporal structure of discovered patterns. The ability to infer latent CP patterns is a vital and foundational component for CP analysis. Our investigations enable both richer representation and more accurate extractions of different aspects of treatment behaviors in CPs, and hence the outcomes of out study can be potentially valuable to CP analysis and redesign.

The remainder of the paper is organized as follows. Section 2 presents preliminary knowledge of the proposed approach. Section 3 describes the proposed CPM for discovering underlying CP patterns from clinical event logs. Section 4 carefully presents our experimental results. Section 5 discusses the results obtained. Related work is outlined in Section 6. Finally, some conclusions are given in Section 7.

2. Preliminaries

2.1. Representation of a patient trace

The objective of this study is to discover latent CP patterns from clinical event logs. In particular, the proposed approach assumes that it is possible to record clinical events sequentially such that each event refers to a treatment activity (i.e., a well-defined step in a CP). Furthermore, additional information such that the occurring time stamp of the event is used in this study. To explain the kind of input needed for the proposed approach, we first define the concept of a clinical event.

Definition 1. (Clinical event)[2]. Let A be a finite set of event identifiers (clinical terms describing treatment activities), and T the time domain (set of time stamp primitives). A **clinical event** e is a pair $e = (a,t)$ where $a \in A$ and $t \in T$. Formally, we use $e{\cdot}a$ and $e{\cdot}t$ to denote the activity type, and the occurring time stamp of a clinical event, respectively. We denote by $E \subseteq A \times T$ the set of all valid clinical events of a particular domain.

Note that clinical events could be characterized by various properties, e.g., an event has an occurring time stamp, it corresponds to an activity type, and has associated costs, etc. We do not impose a specific set of properties, however, given the focus of this paper, we assume that the activity type and occurring time stamp of the event are present. Thus, unfolding in time stamp and activity type, a clinical event e is a treatment activity–time stamp pair $\langle \text{activity}, \text{time stamp} \rangle$ that the activity has been occurred at a specific time point. For example, let '(Admission, day 1)' is a clinical event indicating that a patient was in admission on the first day in his or her length of stay; likewise '(PCI, Day 4)' means the patient was performed a percutaneous coronary intervention on the 4th day in his/her length of stay. For simplicity, the time stamps of these event examples are integer values, however it could be presented in a date-format time stamp.

Definition 2. (Patient trace). Let E be the domain of clinical events (i.e., the number of unique clinical events in the collection). A **patient trace** is a non-empty sequence of clinical events performed on a particular patient, i.e., $\sigma = \langle e_1, e_2, \ldots, e_n \rangle$, where $e_i \in E$ $(1 \leqslant i \leqslant n)$ is a particular clinical event.

In general, a patient trace consists of different categories of clinical events. For a particular patient trace, certain temporal constraints exist between the events. Table 1 depicts an example of eight patient traces. The corresponding meanings of treatment activities listed in Table 1 can be found in Table 4. We will continue to use this example in the rest of the paper to illustrate various points.

From the practical point of view, we make the assumption that treatment behaviors in the executions of CPs are correctly recorded in a clinical event log. Information in the log is expected to (1) refer to clinical events of specific patient traces, and (2) the activity type and occurring time stamp of clinical events. The formal definition of a clinical event log is as follows:

Definition 3. (Clinical event log). Let $\mathcal{C}$ be the set of all possible patient traces. A **clinical event log** $\mathcal{L}$ is a set of patient traces, i.e., $\mathcal{L} \subseteq \mathcal{C}$.

[2] Some clinical events might have a duration, i.e., they are conducted not at a specific time stamp, but over a time period. However, such a clinical event can be assumed to consist of a pair of sub clinical activities, i.e., a start event and an end event, which correspond to a start event and an end event, respectively. In this study, we assume that clinical events are time point events, and intervals are represented by starting and ending time point events.