Knowledge-Based Systems 84 (2015) 203–213

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Supporting healthcare management decisions via robust clustering of event logs

Pavlos Delias [a], Michael Doumpos [b], Evangelos Grigoroudis [b,*], Panagiotis Manolitzas [b], Nikolaos Matsatsinis [b]

[a] Eastern Macedonia and Thrace Institute of Technology, Agios Loukas, 65404 Kavala, Greece
[b] Technical University of Crete, University Campus, Kounoupidiana, 73100 Chania, Greece

ARTICLE INFO

Article history:
Received 8 August 2014
Received in revised form 22 February 2015
Accepted 13 April 2015
Available online 18 April 2015

Keywords:
Process mining
Clustering
Robustness
Knowledge discovery
Healthcare management

ABSTRACT

Business processes constitute an essential asset of organizations while the related process models help to better comprehend the process and therefore to enable effective process analysis or redesign. However, there are several working environments where flows are particularly flexible (e.g., healthcare, customer service) and process models are either very hard to get created, or they fail to reflect reality. The aim of this paper is to support decision-making by providing comprehensible process models in the case of such flexible environments. Following a process mining approach, we propose a methodology to cluster customers' flows and produce effective summarizations. We propose a novel method to create a similarity metric that is efficient in downgrading the effect of noise and outliers. We use a spectral technique that emphasizes the robustness of the estimated groups, therefore it provides process analysts with clearer process maps. The proposed method is applied to a real case of a healthcare institution delivering valuable insights and showing compelling performance in terms of process models' complexity and density.

1. Introduction

Business processes are valuable assets of every organization. They control the revenue potential as much as they shape the cost profile of an organization. Processes directly affect the attractiveness of products and services as perceived by the market and they define the ability of organizations to adapt to new circumstances [10]. Therefore, it is no surprise that organizations strive to model, revise, and optimize their internal business processes, as well as the processes shared with other organizations.

In working environments with strong behavioural diversity (i.e., environments where deviations in the process control flows are common), business models are usually ambiguous [13]. In such environments, the problem concerning business process awareness can be defined as follows: are there any dominant patterns of process behaviour? Is it possible to identify groups of cases with similar behaviours? The objective of this paper is to propose a method that delivers compact and comprehensive synopses of flexible behaviours, keeping in mind the end goal to best support their analysis and improvement.

As an example, in this paper we consider a case study involving the clinical pathways of patients in a hospital, where there is a diverse set of paths followed depending on the peculiarities of each patient. The resulting complex behaviour of the business processes in such an environment can be observed through the trace that every patient leaves. On that account, a process mining perspective is followed in this study. The idea of process mining is to discover, monitor and improve real processes by extracting knowledge from event logs, which are readily available in business information systems [37]. Event logs may store additional information about events (like the timestamp and the resource performing the activity). In other words, each case is leaving a trace, which corresponds to the observed behaviour.

When it comes to clinical pathways analysis, process mining techniques face a critical challenge: Patients routes vary significantly and in order to deliver comprehensive models, the event log should be someway summarized [34]. The authors in [17] propose a horizontal summarization, by partitioning the event log into time intervals. In [16], the authors exploit a rich dataset of patients' traces to summarize the clinical pathways based on a behavioural topic analysis. Indeed, as the authors in [19] discuss, the integration of medical knowledge can significantly improve the comprehensibility of the results.

* Corresponding author. Tel.: +30 28210 37346; fax: +30 282106940.
E-mail address: vangelis@ergasya.tuc.gr (E. Grigoroudis).

http://dx.doi.org/10.1016/j.knosys.2015.04.012
0950-7051/

Often, however, such medical knowledge is not available (e.g., relevant data are not recorded, or data are considered too sensitive to be provided, or even medical experts are not available for the process analysis project). In this paper, we focus on such cases; therefore, we follow a trace clustering approach that mostly relies on the control-flow features of the cases.

Trace clustering aims at discovering clusters with related behaviours. However, considering the set of traces in the event log all at once often leads to ambiguities because the event log contains traces of cases that may refer to very different behaviours (i.e., potentially unique or infrequent cases). By identifying clusters of diverse traces, process discovery techniques could be connected to subsets of behaviours and subsequently deliver more clear, coherent and comprehensible process models.

In this framework, this study contributes by proposing a method that relegates the effect of infrequent behaviours (without ignoring them) and eventually provides effective summarizations of the event log. This is achieved through clustering the traces using a more stable similarity metric. The stability of the metric is reached by introducing the concept of neighbourhood. This addition allows promoting any prevalent patterns, while it reduces the impact of isolated cases to the clusters' formation. In this way the proposed methodology provides compact information and meaningful insights to managers as it facilitates the derivation of a simple interpretation of a complex business process, thus allowing process stakeholders to communicate on an evidence-based ground.

The rest of the paper is organized as follows. The next section provides a brief overview of related works. In Section 3, we describe the case study of a public hospital and the proposed methodology. The approach is analytically presented in Section 4, while the obtained results are discussed in Section 5. Finally, a short discussion concludes the paper.

2. Related works

Flow variability in healthcare processes arises due to the highly customized medical guidelines that describe how patients are treated. Furthermore, it is possible that process analysts and stakeholders do not actually need a complete process model, but just an understanding of a dominant behavioural pattern. In cases where a process is expected to be realized over instances with very different behaviour, discovering a single model would seldom provide clear answers, since the generated models would be complex and confusing (i.e., "spaghetti" models as in [35], p. 301). Clustering different behaviours and discovering a process model per cluster has been identified as an effective solution [11].

An initial and influential approach, presented in [33], proposes the creation of feature vectors for each trace followed by the application of common clustering techniques. Features could be bag-of-activities, transitions, resources, case attributes, etc., while clustering techniques include *k*-means, agglomerative hierarchical clustering, and self-organizing maps. That work introduces the concept of "profiles" for traces, which allows for context information to be considered. However, the stability of the results is not discussed. An ordinary clustering technique (e.g., agglomerative hierarchical clustering) is also used in [18]. In this case, traces are evaluated for their similarities by the activities and the transitions vector. This similarity metric is simple, yet quite straightforward to infer control-flow similarities. While hierarchical clustering is effective in showing how different traces differ from each other, this form of clustering has its disadvantages. The primary disadvantage is that hierarchical clustering is only effective at splitting small amounts of data. When the event log is small, patterns and relationships between clusters are easily discernable. As the event log grows, so does the dendrogram, and this usually results in the loss of information. Besides, all determinations are strictly based on local decisions and a single pass of analysis.

The authors in [41] try to resolve spaghetti models through sequence clustering, i.e., identify frequent sequences of activities through a Markov chain representation. The proposed method could support the post-processing of cluster models (e.g., by discarding infrequent elements). However, the applied algorithm could result in multiple cluster solutions. For instance, the applied migrating-drifting means approach makes the final cluster solution dependent, to some extent, on the order in which the traces are considered for relocation.

Another approach is to use syntactic techniques which operate on the whole sequence "as-is" by applying string distance metrics such as the Levenshtein distance and the generic edit distance, in conjunction with standard clustering techniques [2,3]. A distinctive feature of this approach is that instead of assuming the causes that could explain the variation in process instances (e.g., due to different time periods as noted in [23]) – a task that requires intensive domain knowledge – clusters are created based on a simple similarity metric and variability causes are induced a posteriori (that is, we gain knowledge about the domain). An additional contribution is that the whole method is centred on the robustness of the final solution.

Many approaches, from the area of management and information technology, can be adopted by a healthcare organization in order to optimize its efficiency and effectiveness and to be competitive [22]. The authors in [8] provide a brief overview of business intelligence techniques applied to healthcare services. Moreover, data mining approaches can uncover new biomedical and healthcare knowledge for clinical and administrative decision making as well as generate scientific hypotheses from large experimental data [46]. Should the focus of the research is in discovering rules for temporal patterns (and not process models like in this work), several methods based on local patterns mining can be employed. Such rules are extracted as sequence patterns [7,12,29]. Another approach is to exploit temporal probabilistic models to model healthcare problems. In this category, Bayesian networks are the most visible technique [42]. An additional potential is to exploit temporal data of healthcare services to build predictive models. To this end, different learning algorithms have been applied. The focus of these works is on building predictive data mining models with temporal data (see [1]) using supervised or semi-supervised techniques, like positive-unlabelled learning [15]. However, data mining approaches are data-centric and not process centric [36]. Thus, their output is not directly related to the process mining approach proposed in this work.

Concentrating on process mining techniques, a visible work is that in [30], where the authors developed a methodology for the application of process mining techniques that leads to the identification of regular behaviour, process variants and exceptional medical cases. An additional use of process mining is to check for conformance (process stakeholders can match the assumed process model with the real one – derived from discovery in the event logs) and check if medical standards or administrative guidelines are followed.

3. Problem description

3.1. Introduction to case study

The hospital under consideration in this study is situated in the city of Chania, Greece. It was established in 2000 and has a capacity of 465 beds and 36 operating departments. The hospital is a general public health unit, providing first and secondary degree health