Expert Systems with Applications 41 (2014) 4648–4657

Contents lists available at ScienceDirect

Expert Systems with Applications

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

A dynamic understanding of customer behavior processes based on clustering and sequence mining

Alex Seret [a,*], Seppe K.L.M. vanden Broucke [a], Bart Baesens [a,b,c], Jan Vanthienen [a]

[a] Department of Decision Sciences and Information Management, KU Leuven, Naamsestraat 69, B-3000 Leuven, Belgium
[b] School of Management, University of Southampton, Highfield, Southampton SO17 1BJ, United Kingdom
[c] Vlerick, Leuven-Gent Management School, Reep 1, B-9000 Gent, Belgium

ARTICLE INFO

Keywords:
Clustering
Sequence mining
Business knowledge
Behavior process
Trajectories
Direct marketing

ABSTRACT

In this paper, a novel approach towards enabling the exploratory understanding of the dynamics inherent in the capture of customers' data at different points in time is outlined. The proposed methodology combines state-of-art data mining clustering techniques with a tuned sequence mining method to discover prominent customer behavior trajectories in data bases, which — when combined — represent the "behavior process" as it is followed by particular groups of customers. The framework is applied to a real-life case of an event organizer; it is shown how behavior trajectories can help to explain consumer decisions and to improve business processes that are influenced by customer actions.

1. Introduction

Various data mining techniques have been proven to be a valuable approach in the quest for knowledge discovery in data from an exploratory point of view. Clustering techniques, for instance, combined with strong visualization techniques, allow analysts to get fast insights into the data they are confronted with. For these reasons, techniques such as *k*-means clustering and self-organizing maps have been widely and successfully applied in practice and extensively discussed in the literature (Kohonen, 2001).

When executed at one specific moment in time, however, as it often happens, the aforementioned techniques offer a static picture describing the composition of the data set at hand based on certain patterns derived from the attributes characterizing the instances in this data set (see e.g. Zorrilla & Garcia-Saiz, 2013; Li, Huang, Li, & Zhang, 2011; Carlei, Marra, & Pozzi, 2012). It would, however, be of great interest for the analyst to be able to understand the dynamics associated with the items represented in the data base, hence recording a "movie" of the data set instead of static pictures at specific points in time. This concept is denoted "trajectory" or "customer behavior trajectory" in the remainder of this paper. By describing an object using different attributes, it is possible to obtain a state which describes this object. When repeating this description at different points in time, a sequence of states, or trajectory, is obtained and can be analyzed. In a case where the object of interest is a customer, different attributes linked to her behavior can be captured at a specific point in time and will provide a description of the state of this customer, also called customer behavior. By repeating this description, a customer behavior trajectory is obtained.

In this paper, which is a journal extension of Seret, vanden Broucke, Baesens, and Vanthienen (in press), an approach enabling the exploratory understanding of such dynamics inherent in the capture of customers' data at different points in time is proposed. The contribution of this paper is twofold. First, a general methodology is proposed and offers a comprehensible way to analyze movements in high dimensional spaces using unsupervised methods and visualization. Although multiple researchers are working on dynamic clustering or trajectory mining, few of them are really interested in the comprehensibility of the results for practitioners, which is one of the main research motivations of this work. Broadly summarized, our novel approach is based on a two-step clustering approach, incorporating both self-organizing maps and *k*-means that will generate coordinate sequences used as input for a sequence mining technique. The proposed methodology combines these methods to discover prominent customer behavior trajectories in data bases, which together help analysts to understand the behavior process as it is followed by particular groups of customers. Second, the methodology is applied in order to answer a complex business question in a real-life ticketing context. From a business perspective, understanding the dynamics of customer behaviors is a logical next step for companies applying segmentation techniques to understand their customers since, by definition,

* Corresponding author. Tel.: +32 16 326881.
E-mail address: alex.seret@kuleuven.be (A. Seret).

http://dx.doi.org/10.1016/j.eswa.2014.01.022
0957-4174/

they may not stay indefinitely in the same segments. Capturing these movements becomes then a crucial objective which can only be achieved if comprehensible techniques are proposed. With this in mind, the step-wise visual approach proposed in this work aims not only at identifying the movements but also at reporting them in a way comprehensible for end-users. These different considerations show the relevance of this work for both researchers and practitioners. Moreover, thanks to the general methodology proposed in Section 3.4, the experiments of Section 4 can be easily repeated in other contexts.

The remainder of the paper is structured as follows, in Section 2, an overview of related work is provided. Next, in Section 3, the different techniques and approaches used in the remainder of the paper are introduced from a theoretical perspective. In Section 4, an application using real-life data from the concert industry is proposed and illustrates how the different concepts and techniques can be combined in order to answer advanced business questions. Section 5 concludes the paper.

2. Related work

Some related works following the idea of applying a dynamic approach towards exploratory data mining — and hence, clustering — have been introduced in different works. Some relevant examples are introduced in what follows. In Skupin and Hagelman (2005), the authors propose an approach for the spatialization of multi-temporal, multi-dimensional trajectories using the self-organizing map method and provide the reader with different visualization techniques combined with traditional GIS data structures in order to visualize demographic trajectories. In Sperandio and Coelho (2006), Markov models are build using a self-organizing map to represent the different states of a process. The methodology is then used as a tool for reliability assessment in a power system network. Finally, in a recent work, Chen, Ribeiro, Vieira, and Chen (2013) proposes a methodology for the clustering and visualization of bankruptcy trajectory using self-organizing map. In this approach, two self-organizing maps are trained. The first network uses vectors characterizing banks as input, offering coordinates used to generate trajectories. A second network is then trained to cluster the obtained trajectories and visualize them. The main considerations differencing the approach of Chen et al. (2013) from the one proposed in this paper are the use (in this paper) of sequence mining techniques to generate the frequent trajectories, the introduction of business knowledge to guide the clustering algorithm, the removal of repetitions as focus is put on movements rather than the duration of a particular item remaining in a certain cluster and the introduction of statistical descriptions of the identified movements. As mentioned in the above, our proposed methodology incorporates a modified sequence mining procedure similar as described in Agrawal, Mannila, Srikant, Toivonen, and Verkamo (1996). In recent years, a new research field denoted as "process mining" has sprung up, aiming to extract valuable knowledge from event based data repositories ("event logs"), including the extraction of a high-level business process model, capturing thus also an aggregated, frequency based dynamic view over the given input (see Agrawal, Gunopulos, & Leymann, 1998; Cook & Wolf, 1998; van der Aalst, Weijters, & Maruster, 2004; Rozinat & van der Aalst, 2006). Contrary to this collection of techniques, however, our described approach is data driven using derived state (rather than event) sequences derived from instance level attributes collected over time, meaning that no full event log data is required to utilize the outlined methodology. In addition, state-of-art clustering techniques are applied, allowing for the identification of different customer groups (based on the behavior patterns discovered in the data) behind the same decision outcome (e.g. opting for a subscription).

3. Theoretical approach

In this section, the different techniques supporting our dynamic approach are discussed. References to related work are also described in this section. For the purpose of the application, a two-step clustering approach is considered, leading to clusters that will be used further on to generate trajectories summarized using a sequence mining approach. The following subsections introduce respectively the knowledge-based constrained clustering, the *k*-means algorithm and the generalized sequential pattern algorithm. Section 3.4 integrates the different concepts to formulate the proposed methodology.

3.1. Knowledge-based constrained clustering

In order to incorporate business knowledge in the segmentation exercise, the P-SOM algorithm proposed in Seret, Verbraken, and Baesens (submitted for publication) will be used in Section 4 to contrast the classical SOM algorithm. This algorithm is a modified version of the traditional SOM algorithm providing a mechanism which enables the prioritization of variables. Both SOM and P-SOM algorithms are shortly introduced in this section. The P-SOM being a variant of the SOM algorithm, it is worth spending some time understanding the classical SOM algorithm, which key concepts are discussed in Section 3.1.1, before reading the Section 3.1.2 where the P-SOM itself is introduced.

3.1.1. SOM

Self-organizing maps (SOM) have been widely discussed in the literature and successfully applied in the practice (Kohonen, 1995; Smith & Ng, 2003; Schwartz, Smith, Churilov, Dally, & Weber, 2003; Huysmans, Martens, Baesens, Vanthienen, & Van Gestel, 2006; Seret, Verbraken, Versailles, & Baesens, 2012; Louis, Seret, & Baesens, 2013). Introduced in 1981 by Teuvo Kohonen, this approach based on artificial neural networks (ANNs) aims at summarizing and projecting high dimensional data onto a lower dimensional space. This technique combines advantages of both vector quantization (e.g. the *k*-means algorithm introduced in Section 3.2) and vector projection (e.g. the PCA Jolliffe, 2005) techniques, enabling, inter alia, visual clustering and correlation analysis, outliers detection, noise reduction and supervised learning. Based on an efficient competitive learning approach, neurons are trained in order to capture the structure of the training data set while preserving the topological properties of the input. The first step of the algorithm consists of the initialization of the neurons that will later be trained. The input vectors n_i are then presented randomly to the network until convergence is achieved or a predefined number of iterations are performed. When presenting an input vector n_i to the algorithm, the closest neuron, the best matching unit (BMU), m_c is identified:

$$\|n_i - m_c\| = \min_r(\|n_i - m_r\|). \tag{1}$$

The weights of this BMU are then updated together with the neurons identified as being part of the neighborhood of the BMU, leading to the following learning function:

$$m_r(t+1) = m_r(t) + \alpha(t)h_{cr}(t)[n_i(t) - m_r(t)], \tag{2}$$

with $\alpha(t)$ and $h_{cr}(t)$ being respectively a learning rate and a neighborhood function. The learning rate will impact the magnitude of the update while the neighborhood function will define the set of neurons impacted by the update. Both the learning rate and the neighborhood functions are decreasing during the training in order to obtain a stable solution. In order to evaluate the resulting output, the mean quantization error (MQE) and a topographic function are commonly used (Kohonen, 2001; Pölzlbauer, 2004). On the one