European Journal of Operational Research 234 (2014) 720–730

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Stochastics and Statistics

Mining categorical sequences from data using a hybrid clustering method

Luca De Angelis [a,*], José G. Dias [b]

[a] Department of Statistical Sciences, Alma Mater Studiorum, University of Bologna, Via Belle Arti, 41, 40126 Bologna, Italy
[b] Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit, Portugal

ARTICLE INFO

Article history:
Received 5 January 2012
Accepted 5 November 2013
Available online 23 November 2013

Keywords:
Data mining
Sequential data
Hidden Markov models
Clustering
Categorical data

ABSTRACT

The identification of different dynamics in sequential data has become an every day need in scientific fields such as marketing, bioinformatics, finance, or social sciences. Contrary to cross-sectional or static data, this type of observations (also known as stream data, temporal data, longitudinal data or repeated measures) are more challenging as one has to incorporate data dependency in the clustering process. In this research we focus on clustering categorical sequences. The method proposed here combines model-based and heuristic clustering. In the first step, the categorical sequences are transformed by an extension of the hidden Markov model into a probabilistic space, where a symmetric Kullback–Leibler distance can operate. Then, in the second step, using hierarchical clustering on the matrix of distances, the sequences can be clustered. This paper illustrates the enormous potential of this type of hybrid approach using a synthetic data set as well as the well-known Microsoft dataset with website users search patterns and a survey on job career dynamics.

1. Introduction

The identification of different dynamics in sequential data has become a fundamental step in many research fields. For instance, the analysis of the gene expression dynamics represents a relevant task in the bioinformatics framework (Ramoni, Sebastiani, & Cohen, 2002; Ramoni, Sebastiani, & Kohane, 2002; Tucker, Hoen, Vinciotti, & Liu, 2006), whereas the approximation of the spread of infectious diseases in large populations helps generate efficient dynamic optimization techniques to assist real-time modification of public health interventions (Yaesoubi & Cohen, 2005). In information science, the issue of text categorization which is originally handled using Support Vector Machines can also be tackled evaluating word sequential patterns and thus taking into account the temporal relationships between words and sentences as well (Jaillet, Laurent, & Teisseire, 2006). Other examples can be found in marketing where analyses aim to investigate the consumers' brand choice behavior (Poulsen, 1990), as well as financial studies, where the purpose of researchers is to recognize similar fluctuations among stock markets (Basalto et al., 2007; Ramos, Vermunt, & Dias, 2011), and in modeling manpower systems where both observable and latent sources of dynamic heterogeneity should be accounted for (Guerry, 2011). Furthermore, in social sciences such as economics (Frühwirth-Schnatter & Kaufmann, 2008) and demography (Dias & Willekens, 2005), mining temporal or longitudinal data is essential for obtaining an accurate analysis of phenomena. Examples can also be found in engineering sciences such as reliability analysis (Zhou, Hua, Xu, Chen, & Zhou, 2010), equipment's health status assessment (Dong & He, 2007) and in air traffic flow management where time series data clustering facilitates the mitigation of airport congestion effects (Inniss, 2006).

The challenging task of clustering this type of observations is that one has to incorporate data dependency in the clustering process (Kakizawa, Shumway, & Taniguchi, 1998). Therefore, methods for clustering time series data have recently received a growing attention in both data mining and statistics literature. However, most of the proposals developed so far merely try to modify the existing algorithms for clustering static data in such a way that time series data can be handled (Liao, 2005).

Furthermore, existing literature mainly focus on the analysis of real-valued and discrete-valued data series: the investigation and clustering of the temporal dynamics of sequences of quantitative variables is a preferred topic in time series analysis literature, but procedures for clustering categorical sequential data have only been addressed marginally (Raftery, 1985).

The search for data mining methods for dealing with large data sets has recently increased the interest in sequential data clustering (Ananthanarayana, Murty, & Subramanian, 2001; Keogh & Kasetty, 2003). For instance, the discovery of similar web navigation patterns has become a popular topic in web mining (Vakali,

[*] Corresponding author.
E-mail addresses: l.deangelis@unibo.it (L. De Angelis), jose.dias@iscte.pt (J.G. Dias).

0377-2217/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ejor.2013.11.002

Table 1
Toy example – observed sequences A and B.

Time	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Sequence A	1	1	1	2	2	2	2	2	1	1	1	2	1	2	1	2
Sequence B	1	1	2	2	1	1	2	2	1	1	2	2	1	1	2	2

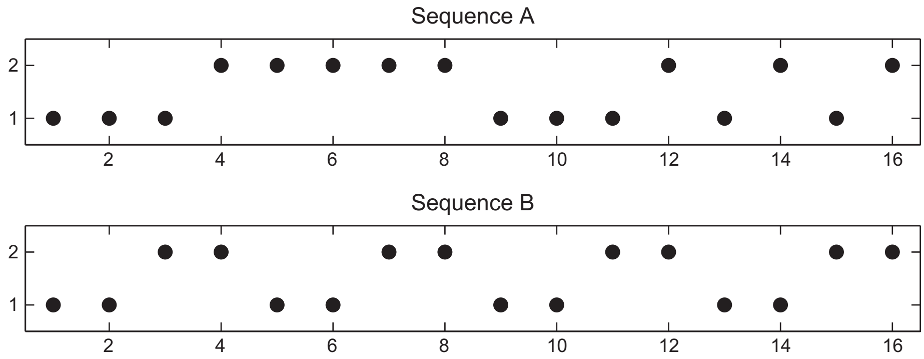

Fig. 1. Toy example – observed sequences A and B.

Pokorny, & Dalamagas, 2004). In this context, Spiliopoulou and Pohle (2001) show that traditional data mining approaches may not be appropriate for detecting different website users search patterns. Thus, many algorithms for clustering web usage patterns have been proposed. For example, approaches which extend the traditional K-means algorithm using the Kullback–Leibler distance as an alternative to the Euclidean distance (Dias & Cortinhal, 2008; Petridou, Koutsonikola, Vakali, & Papadimitriou, 2006); hierarchical pattern-based algorithms adopted for clustering web transactions (Yang & Padmanabhan, 2011); and model-based approaches based on Markov models (Anderson, Domingos, & Weld, 2002; Cadez, Heckerman, Meek, Smyth, & White, 2003; Dias & Vermunt, 2007; Dongshan & Junyi, 2002; Sen & Hansen, 2003) and self-organizing maps (Smith & Ng, 2003).

Ramoni et al. (2002) propose a Bayesian method – BCD algorithm – which models dynamic processes as Markov chains and then hierarchically groups time series according to the Bayes rule by applying an agglomerative clustering procedure. They deal with observed frequencies to compute the transition matrices assuming Markov chains in the first step of their algorithm. In the second step, they compute the average symmetrized Kullback–Leibler distances between (observed) transition matrices and (agglomerative hierarchically) group time series according to the Bayes rule, i.e., evaluating the posterior probabilities of a partition given the data.

We provide a toy example to show the limitations of using observed transitions between categories in the clustering task. Table 1 contains two categorical sequences – A and B – with observed state-space $\{1,2\}$ of length 16. Fig. 1 depicts these sequences. Despite the same original state – both sequences start in category 1 – sequences are remarkably different. Moreover, from a Markov chain point of view both sequences are identical as its sufficient statistics are the same. Indeed, initial states and the matrix of observed transitions between states are identical (Table 2). Thus, based on these input data, any measure of distance between A and B will be null, and sequences are assumed as identical and will belong always to the same cluster.

Liao (2007) deals with continuous multivariate data and he converts them into a discrete-valued univariate time series. Then, he uses the same identical procedure introduced by Ramoni et al. (2002). Bicego, Murino, and Figueiredo (2004) train each sequence by different hidden Markov models. However, this procedure fails in scaling all sequences on the same set of parameter and retrieve comparable parameter estimates.

Table 2
Toy example – observed transitions between states.

Sequence A			Sequence B		
States	1	2	States	1	2
1	4	4	1	4	4
2	3	4	2	3	4

In this paper, we aim to provide a contribution to the methodology of categorical time series clustering which represents an innovative and challenging task. In particular, we propose a hybrid method, which combines the model-based clustering approach provided by an extension of the hidden Markov model (HMM) and a hierarchical clustering procedure. More specifically, we develop a clustering process that incorporates the data dependency observed in the time series and transforms the categorical data set into a probabilistic space, where a symmetric Kullback–Leibler distance can operate. Then, we resort to hierarchical clustering on the matrix of distances in order to cluster the data sequences into homogenous groups according to both their characteristics and dynamics. Moreover, our hybrid method does not require that all sequences have the same length.

The two-step procedure of our novel hybrid method for clustering categorical sequential data and discovering groups characterized by similar trajectories and dynamics allows us to handle two issues which affect many of the aforementioned proposals. The first step of our approach represented by the estimation of the Panel HMM enables us to assume that, conditional on cluster membership, the observations are dependent and, thus, analyzing the sequential structure of the data not only at individual level. On the other hand, the second step, which is derived from a distance-based clustering procedure, addresses the evidence that sequential data dynamics, such as web navigation data, may be non-Markovian in nature (Huberman, Piroli, Pitkow, & Lukose, 1997). Thus, our approach differs from that of Ramoni et al. (2002) in two important ways. In the first step of the clustering procedure, both the BCD algorithm by Ramoni et al. (2002) and our hybrid approach transform the original time series using Markov chains. However, BCD algorithm replaces the categorical sequential series by Markov chains represented by transition probability matrices obtained by computing the *observed* switching between adjacent