Computational Statistics and Data Analysis 93 (2016) 31–45

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Model-based biclustering of clickstream data

Volodymyr Melnykov *

Department of Information Systems, Statistics, and Management Science, The University of Alabama, Tuscaloosa, AL 35487, USA

ARTICLE INFO

Article history:
Received 15 November 2013
Received in revised form 13 September 2014
Accepted 17 September 2014
Available online 28 September 2014

Keywords:
Finite mixture model
Model-based clustering
Biclustering
Clickstream
Model selection

ABSTRACT

Navigation patterns expressed by sequences of visited web-sites or categories can characterize the behavior and habits of users. Such web-page routes taken by individuals are commonly called clickstreams. Clustering clickstream sequences is a recent yet challenging problem with many applications. The main difficulty is related to the fact that one needs to group categorical data sequences rather than vectors and the majority of traditional clustering algorithms are not applicable in this setting. The time-related character of data suggests that dynamic models have a better promise than static ones. Model-based clustering relying on the mixture of first order Markov models will be considered. Since the number of distinct web-pages, and therefore the number of states in a Markov process, can be very high, such a mixture model involves a large number of parameters. Thus, grouping states by their similarity to reduce the number of parameters in the model is also proposed. Then, states are clustered along with users providing a biclustering framework. The developed methodology is illustrated on synthetic and real datasets with good results.

1. Introduction

Clustering sequences of categorical data is a challenging problem that recently attracted attention of researchers. One particular area of application in which this framework occurs is grouping Internet users by their navigation patterns (Banerjee and Ghosh, 2000; Inbarani and Thangavel, 2009; Liu, 2011). Knowing routes taken by customers is important, for example, for major web-based stores as it allows grouping users according to their preferences. Then, the store can advertise to its visitors specific products that other customers with similar navigation behavior eventually acquired. Such user's navigation route is commonly called a clickstream.

In real life, the total number of categories can be low or high. Clickstream sequences associated with users can include very few or numerous categories each of which can be repeated multiple times. Also, the lengths of sequences can vary from very short to extremely long. As remarked in Banerjee and Ghosh (2001), the main difficulty in clustering this type of data is that observations are available in the form of categorical data sequences rather than in the form of vectors, for which cluster analysis techniques are well-established. The authors proposed clustering users based on the longest common subsequence in their clickstreams. As pointed out in Cadez et al. (2003), another major problem is related to the dataset size. Since the number of web-pages visited and especially the number of users can be extremely high, many clustering methods (*e.g.*, hierarchical algorithms) become restrictive. The authors also indicated that a model reflecting the dynamic nature of the problem should be more appropriate than a static one. In light of these remarks, a mixture of first-order Markov models was employed in Cadez et al. (2003). This mixture provides a flexible instrument for modeling dynamic clickstream behavior. The authors based their model on the set of multinomial distributions. However, this is not quite correct as the order of visited web-pages matters and the multinomial coefficient should be abandoned. The above-mentioned paper is the only one in

* Correspondence to: The University of Alabama, Alston Hall 346, Tuscaloosa, AL 35487, USA. Tel.: +1 2053486292.
E-mail address: vmelnykov@ua.edu.

http://dx.doi.org/10.1016/j.csda.2014.09.016
0167-9473/

the clickstream literature that attempts model-based clustering. Some problems, however, might occur when there are numerous states involved in the Markov model. Then, the number of model parameters can become prohibitively high. This is a serious limitation of the mixture-modeling approach that has to be addressed in order to make the proposed technique practical. One can note that there might be many web-pages or web-categories that do not differ considerably. For example, visitors of web-sites containing information about camcorders of two different brands can behave quite similarly. Even visitors of web-pages devoted to camcorders and cameras can be similar in many situations. This suggests combining some states into more general categories such as "camcorders" or "electronics". Then, this translates our clustering problem into a problem of *biclustering*, in which, in addition to objects, features also have to be grouped according to their similarities. In the considered problem setting, this implies that, in addition to clustering web-users, web-pages or web-categories also have to be grouped.

Biclustering ideas have been especially popular in biological applications such as the analysis of gene expression (Li et al., 2012; Madeira and Oliveira, 2004). Another biclustering application arises in text mining with data being summarized in the form of a matrix with rows representing documents and columns representing particular words (Bisson and Hussain, 2008; Dhillon, 2001). The majority of methods focus on searching for the optimal decomposition of matrix blocks. Recently, a visual approach for exploring groups of clickstream data was proposed in Wei et al. (2012). The authors developed an interactive mechanism incorporating self-organizing maps representing a class of neural network models. The first order Markov model was employed to represent user behavior. Perhaps, the main criticism of the developed approach is related to the projection of clickstreams into a two-dimensional space fulfilled by self-organizing maps. While it is easier and more intuitive to work with low dimensions, some important information can be lost. Other work in this area that is worth mentioning includes Montgomery et al. (2004) and Ypma and Heskes (2003). Many biclustering algorithms, however, are problem-specific which restricts their application to a broader domain of problems. A different setting that resembles biclustering and whose ideas are closely related to it is the selection of variables important for clustering (Maugis et al., 2009a,b; Raftery and Dean, 2006). The underlying idea is that there are redundant or noise variables which make the task of grouping observations more challenging. Therefore, the goal is to find the set of variables that provide the best clustering solution. This search for data columns has to be conducted along with grouping observations which closely resembles biclustering.

We approach the clickstream clustering and biclustering problems by means of finite mixture models and model-based clustering (Fraley and Raftery, 1998, 2002; McLachlan and Peel, 2000; Maitra and Melnykov, 2010). The main idea of this technique is to fit data with a collection of distributions in such a way that every mixture component adequately models a particular data group. It can be unrealistic in many cases as several distributions can be required to describe a cluster. In this case, further data and model manipulations, such as merging mixture components, might be necessary (Baudry et al., 2010; Hennig, 2010; Melnykov, 2013a). A recent book by Govaert and Nadif (2013) is an excellent resource on existing biclustering approaches in the model-based framework.

Section 2 discusses basic finite mixture modeling concepts, provides a comprehensive derivation of the model-based clustering approach, develops a variability assessment procedure, and extends the proposed methodology by introducing biclustering ideas relying on model selection procedures. Section 3 evaluates the performance of the developed procedure by means of a rigorous simulation study. Section 4 is devoted to the application of the proposed technique to a real-life dataset. The paper concludes with a brief discussion in Section 5.

2. Methodology

2.1. Mixture modeling and model-based clustering

Let $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_n$ represent a random sample of T-variate observations following a probability distribution function of the form

$$f(\boldsymbol{y}|\boldsymbol{\vartheta}) = \sum_{k=1}^{K} \tau_k f_k(\boldsymbol{y}|\boldsymbol{\vartheta}_k), \tag{1}$$

where $f_k(\boldsymbol{y}|\boldsymbol{\vartheta}_k)$ represents the kth mixture component of known functional form with the parameter vector $\boldsymbol{\vartheta}_k$, τ_k is the kth mixing proportion also known as prior probability, subject to two restrictions: $0 < \tau_k \leq 1$ and $\sum_{k=1}^{K} \tau_k = 1$, K is the total number of mixed distributions, and $\boldsymbol{\vartheta}$ denotes the entire parameter vector $\boldsymbol{\vartheta} = (\tau_1, \tau_2, \ldots, \tau_{K-1}, \boldsymbol{\vartheta}_1^{tr}, \boldsymbol{\vartheta}_2^{tr}, \ldots, \boldsymbol{\vartheta}_K^{tr})^{tr}$ with tr representing the transpose operation. A probability distribution following the form (1) is called a finite mixture model. The traditional method of estimating $\boldsymbol{\vartheta}$ in this framework is maximum likelihood estimation fulfilled by means of the expectation–maximization (EM) algorithm (Dempster et al., 1977). The EM algorithm consists of two steps, called E-step (expectation) and M-step (maximization), that need to be repeated iteratively until the convergence is reached. At the E-step, the conditional expectation of the complete-data log likelihood function given observed data has to be found. This expectation, commonly referred to as the Q-function, must be maximized with respect to the parameter vector $\boldsymbol{\vartheta}$ at the M-step. The stopping criterion most commonly used in finite mixture modeling is based on the relative change in log likelihood values obtained from two consecutive iterations b and $b-1$: $(\ell(\boldsymbol{\vartheta}^{(b)}) - \ell(\boldsymbol{\vartheta}^{(b-1)}))/|\ell(\boldsymbol{\vartheta}^{(b)})|$. The EM algorithm should be terminated when the relative change becomes smaller than some pre-specified tolerance level ϵ. Upon convergence, say at the step B, the maximum likelihood estimate (MLE) $\hat{\boldsymbol{\vartheta}} = \boldsymbol{\vartheta}^{(B)}$ and corresponding posterior probabilities