Econometrics and Statistics 000 (2017) 1–19

Contents lists available at ScienceDirect

Econometrics and Statistics

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

Evolutionary clustering for categorical data using parametric links among multinomial mixture models

Md. Abul Hasnat[a], Julien Velcin[a], Stephane Bonnevay[b], Julien Jacques[a,*]

[a] Université de Lyon, Université Lyon 2 & ERIC, France
[b] Université de Lyon, Université Lyon 1 & ERIC, France

ARTICLE INFO

Article history:
Received 22 March 2016
Revised 6 March 2017
Accepted 6 March 2017
Available online xxx

Keywords:
Evolutionary clustering
Multinomial distribution
Mixture model
Model-based clustering
Twitter data

ABSTRACT

A novel evolutionary clustering method for temporal categorical data based on parametric links among the Multinomial mixture models is proposed. Besides clustering, the main goal is to interpret the evolution of clusters over time. To this aim, first the formulation of a generalized model that establishes parametric links among two Multinomial mixture models is proposed. Afterward, different parametric sub-models are defined in order to model the typical evolution of the clustering structure. Model selection criteria allow to select the best sub-model and thus to guess the clustering evolution. For the experiments, the proposed method is first evaluated with synthetic temporal data. Next, it is applied to analyze the annotated social media data. Results show that the proposed method is better than the state-of-the-art based on the common evaluation metrics. Additionally, it can provide interpretation about the temporal evolution of the clusters.

1. Introduction

In the recent years, the social media plays a significant role in many aspects of our daily activity. There exist numerous popular social media, such as Twitter or Facebook, where the users often provide their opinions about particular entity, e.g., persons (politician, actor), products consumed in the daily life, etc. A common method to analyze such data is – first use a clustering method to group the users/opinions, and then investigate each group independently. An important property of these data is that they may change *over time* due to the changes of attributes, and appearance/disappearance of users. Moreover, users may change their opinion about the targeted entity.

An ordinary clustering method is unlikely to adapt with such temporal dynamics of the data because it does not consider any relevant information, such as history and temporal effects. The notion of evolutionary clustering (Chakrabarti et al., 2006) appears in such situation, where the method should be specialized in clustering temporal data by taking care of the historic information and current data altogether. Numerous methods exist, which address these issues appropriately and cluster temporal data. These methods are based on different strategies, such as spectral clustering (Chi et al., 2009; Xu et al., 2014) and probabilistic generative model (Blei and Lafferty, 2006; Xu et al., 2012; Kim et al., 2015). However, it remains an important issue – how to interpret the evolution of the clusters. In this research, we are motivated by this issue and propose a novel method based on the Multinomial mixture model (Bishop et al., 2006) to cluster the temporal data

[*] Corresponding author at: Laboratoire ERIC, 5 av. Mendès France, 69676 Bron Cedex, France.
E-mail addresses: mhasnat@gmail.com (Md.A. Hasnat), julien.velcin@univ-lyon2.fr (J. Velcin), stephane.bonnevay@univ-lyon1.fr (S. Bonnevay), julien.jacques@univ-lyon2.fr (J. Jacques).

http://dx.doi.org/10.1016/j.ecosta.2017.03.004
2452-3062/© 2017 EcoSta Econometrics and Statistics. Published by Elsevier B.V. All rights reserved.

as well as interpret the evolution of the clusters through some prior belief. Therefore, we propose a novel method which simultaneously performs evolutionary clustering and interpret the evolution.

Multinomial Mixture (MM) model based clustering strategy is a popular method for clustering discrete data (Meilă and Heckerman, 2001; Agresti, 2002). Most recently, it has been exploited to perform evolutionary clustering (Kim et al., 2015). In this research, we consider MM as the core model for the data and propose an evolutionary clustering method by deriving appropriate link between the parameters of MM at different time.

Parametric link among probability distributions has been used in the context of transfer learning (Biernacki et al., 2002; Beninel et al., 2012), where the goal is to adapt a clustering model from a source population to a target one. In the context of continuous features, (Biernacki et al., 2002) proposed a parametric link between the Normal distributions. Jacques and Biernacki (2010) extended it for the binary features using the Bernoulli distribution. However, no such formulation exists for the Multinomial distribution. Moreover, such parametric link-based methods are never considered in the context of evolutionary clustering. We are motivated from both of these issues and propose a clustering method that exploits the links among the parameters of the Multinomial distributions to analyze the temporal/evolutionary data.

Our overall contribution in this research is to propose a novel evolutionary clustering method based on the Multinomial mixture model. The highlights of our contributions include: (a) propose a formulation for parametric link among the Multinomial distributions; (b) develop a novel evolutionary clustering method by exploiting the link parameters and (c) provide interpretation of the link parameters to describe cluster evolution. First, we use synthetic data to evaluate and compare the proposed method w.r.t. the state-of-the-art methods. Next, we apply it to analyze the temporal dynamics of social media data obtained from the *ImagiWeb* project (Velcin et al., 2014). Results in Section 4 show that the proposed method is better than the state-of-the-art methods.

In the rest of the paper, we provide related background in Section 2, describe our proposed method in Section 3, present the experimental results and observations in Section 4 and finally draw conclusions in Section 5.

2. Background and related work

Evolutionary Clustering (ECL), also called *clustering over time*, aims to cluster the data that dynamically evolves over time (Chakrabarti et al., 2006). ECL methods cluster the data by considering the temporal smoothness to reflect the long-term trends of the data while being robust to the short-term variations. The demand and applications of these methods are increasing rapidly in various domains. They have been successfully applied to analyze news (Xu et al., 2012), social media (Kim et al., 2015), stock price (Xu et al., 2014), photo-tag pairs (Chakrabarti et al., 2006) and documents (Blei and Lafferty, 2006).

Temporal/evolutionary data clustering has been addressed from several viewpoints in the literature, which naturally raises several task-specific notions about ECL. A distinction among them can be as follows: (1) clustering; (2) monitoring and (3) interpreting. In the following paragraphs, we review relevant work based on this distinction.

Following the definition of Chakrabarti et al. (2006), the ECL method clusters data by considering the historic and current information. Based on this definition, we do not consider the methods which do not take into account the historic information. Besides, in order to limit our focus on the parametric methods, we do not consider the methods from non-parametric Bayesian based approaches (Xu et al., 2008; Dubey et al., 2013; Kharratzadeh et al., 2015).

Numerous ECL methods have been proposed in the literature. Chakrabarti et al. (2006) provided a generic framework and proposed different versions with the k-means and hierarchical clustering. It is based on optimizing a global cost function, which incorporates the snapshot (static clustering) quality and history cost (temporal smoothness). Chi et al. (2009) proposed two methods based on spectral clustering. They added different terms within the cost functions to regularize the temporal smoothness. Xu et al. (2014) recently proposed AFFECT, which performs adaptive evolutionary clustering by estimating an optimal smoothing parameter. It is extended with several static methods, such as k-means, hierarchical and spectral. A common property of these methods is that they are specialized for continuous data. Therefore, they may not be an appropriate choice for the categorical data, which is our concern in this research.

Dynamic Topic Model (DTM) is a well-known method for analyzing temporal categorical data (Blei and Lafferty, 2006). It extends the popular topic modeling method called Latent Dirichlet Allocation (LDA) (Blei et al., 2003). It uses Dirichlet prior based smoothing, which sometime over-smooth the data. As a consequence, it may cluster the data samples with non-co-occurring features in the same group (Kim et al., 2015). This causes DTM to underperform when clustering the classical non-textual temporal categorical data. Recently, Kim et al. (2015) address this issue and proposed Temporal Multinomial Mixture (TMM). TMM extends the classical Multinomial mixture (MM) model by incorporating temporal dependency into the relation between the data of current time epoch and the clusters of the previous time epoch. Indeed, TMM is more related to our proposed approach as we aim to establish parametric link among MMs at different time epochs. Unfortunately, both DTM and TMM are unable to detect and interpret cluster evolution, which is one of the main foci of this research.

Evolution *monitoring* (Spiliopoulou et al., 2006) tracks the clusters evolution by identifying the birth, death, split, merge and survival of clusters at different time. An external clustering method is first used at each time, e.g., Spiliopoulou et al. (2006) and Oliveira and Gama (2010) used the k-means method, whereas Lamirel (2012) used the neural clustering method. Afterward, the mapping among the clusters at different time is examined based on several heuristics. A different method, called label-based diachronic approach (Lamirel, 2012), exploits the MultiView Data Analysis technique among the cluster labels at different time. This approach constructs heuristics from features for monitoring cluster evolution. Our approach is