Computational Statistics and Data Analysis 103 (2016) 206–228

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

A variational Expectation–Maximization algorithm for temporal data clustering

Hani El Assaad [a,*], Allou Samé [a], Gérard Govaert [b], Patrice Aknin [a,c]

[a] Université Paris-Est, IFSTTAR, GRETTIA, F-77447 Champs-Sur-Marne, France
[b] Université de technologie de Compiègne, UMR CNRS 7253 Heudiasyc, F-60205 Compiègne, France
[c] SNCF - Research & Innovation, F-75012 Paris, France

ARTICLE INFO

Article history:
Received 26 February 2015
Received in revised form 11 May 2016
Accepted 12 May 2016
Available online 20 May 2016

Keywords:
Temporal data clustering
Dynamic latent variable model
Mixture model
EM algorithm
Kalman filter
Clustering
Maximum likelihood
Variational approximation

ABSTRACT

The problem of temporal data clustering is addressed using a dynamic Gaussian mixture model. In addition to the missing clusters used in the classical Gaussian mixture model, the proposed approach assumes that the means of the Gaussian densities are latent variables distributed according to random walks. The parameters of the proposed algorithm are estimated by the maximum likelihood approach. However, the EM algorithm cannot be applied directly due to the complex structure of the model, and some approximations are required. Using a variational approximation, an algorithm called VEM-DyMix is proposed to estimate the parameters of the proposed model. Using simulated data, the ability of the proposed approach to accurately estimate the parameters is demonstrated. VEM-DyMix outperforms, in terms of clustering and estimation accuracy, other state-of-the-art algorithms. The experiments performed on real world data from two fields of application (railway condition monitoring and object tracking from videos) show the strong potential of the proposed algorithms.

Contents

* Correspondence to: French Institute of Science and Technology for Transport development and networks (IFSTTAR), 14-20 Boulevard Newton Cité Descartes, 77447 Champs-Sur-Marne, France. Tel.: +33 18 166 8713.
E-mail address: hani.el-assaad@ifsttar.fr (H. El Assaad).

http://dx.doi.org/10.1016/j.csda.2016.05.007
0167-9473/© 2016 Elsevier B.V. All rights reserved.

1. Introduction

Cluster analysis, which consists in automatically identifying groups into data sets, remains a central issue in many applications including web data mining, marketing, bio-informatics, image segmentation and text mining. The Gaussian mixture model (GMM) (McLachlan and Peel, 2004; Titterington et al., 1985), used conjointly with the Expectation–Maximization (EM) algorithm (Dempster et al., 1977), is now well known to provide powerful clustering solutions. However, some challenges still remain for the processing of non-stationary data.

This study was motivated by the clustering of temporal data acquired on some critical railway components, for characterizing the dynamic of their degradations. Its final objective is to build a decision-aided support for their preventive maintenance. To solve this problem, we propose to automatically extract, from temporal data, clusters whose centers evolve over time.

The general situation, where at each time a set of multivariate observations is acquired, is considered in this article. Fig. 1 shows an example of such temporal data, where we have, for instance, three observations at $t = 1$ and five observations at $t = 2$.

One way to address this specific clustering problem is to assume that the data are distributed according to a Gaussian mixture model whose centers are linear functions of time (DeSarbo and Cron, 1988; Wedel and DeSarbo, 1995). However, a linear evolution of the clusters may turn out to be inefficient for complex nonlinear dynamics. For tracking time-varying spike shapes, Calabrese and Paninski (2011) have proposed a method which consists of maximizing the log-likelihood criterion associated to the classical Gaussian mixture model, penalized by a term that takes into account the temporal evolution of the clusters.

In this work, a dynamic latent variable model dedicated to temporal data clustering is introduced. The frequentist approach was adopted to estimate its parameters. Unfortunately, estimating the parameter of this model by the maximum likelihood approach via the EM algorithm is intractable. Difficulties arise in the E step due to the dynamic structure of the model, which requires integrations over all possible configurations of the hidden variables. Some approximations are therefore required.

Viewing EM as the alternate optimization of an auxiliary function, respectively with respect to the distribution over the latent variables and the parameters (Neal and Hinton, 1998), a variational approximation is proposed in this paper. The idea is to restrict the latter optimization problem to a family of distribution over latent variables, which can be factorized into independent factors. In this case, a lower bound of the log-likelihood criterion is maximized. Variational inference, which was initiated in the mid-1990s, is usually applied to complex models involving missing values or based on latent structures when the direct implementation of standard EM is difficult to be achieved (Jaakkola and Jordan, 1997; Jordan et al., 1998). It has been proved to provide relevant estimates of mixture models in different configurations (Govaert and Nadif, 2008).

Alternative methods can be used to tackle the maximum likelihood problem, such as stochastic versions of EM (see McLachlan and Krishnan, 2008, chap. 6). For instance, the SEM-Gibbs algorithm proposed by Keribin et al. (2010) runs a Gibbs sampler to simulate the unknown labels. However, we opted for a variational approximation, which is computationally more attractive.

The paper is organized as follows. Section 2 briefly reviews the penalized maximum likelihood approach of Calabrese and Paninski (2011). In Section 3, a dynamic model is formalized for clustering temporal data, and a new parameter estimation method based on a variational EM algorithm is presented. An incremental version of the proposed algorithm is formulated in Section 4. Experiments carried out on simulated and real data are presented in Section 5. Finally, conclusions and future works are proposed in Section 6.

The following notations will be used throughout this paper: $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$ denotes the sequence of T observed data to be classified, where $\mathbf{x}_t$ is itself a sub-sample of n_t multivariate observations $(\mathbf{x}_{t1}, \ldots, \mathbf{x}_{tn_t})$, with $\mathbf{x}_{ti} \in \mathbb{R}^d \ \forall i = 1, \ldots, n_t$. The unobserved classes associated to the observations will be denoted by $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_T)$, where $\mathbf{z}_t = (z_{t1}, \ldots, z_{tn_t})$, with $z_{ti} \in \{1 \ldots, K\}$. The means of the Gaussian densities will be denoted as $\boldsymbol{\mu} = (\boldsymbol{\mu}_k^{(t)};\ t = 1, \ldots, T,\ k = 1, \ldots, K)$.

To simplify the notations, the sums and products relative to time, observations at each time and clusters will be subscripted respectively by the letters t, i, k without indicating the limits of variation. So, for instance, the sum $\sum_t$ stands for $\sum_{t=1}^{T}$, the sum $\sum_i$ stands for $\sum_{i=1}^{n_t}$, $\sum_{t,i,k}$ stands for $\sum_{t=1}^{T} \sum_{i=1}^{n_t} \sum_{k=1}^{K}$ and $\prod_{t,i,k}$ stands for $\prod_{t=1}^{T} \prod_{i=1}^{n_t} \prod_{k=1}^{K}$.

2. A penalized likelihood approach for temporal data clustering

This section gives a brief review of the temporal data clustering approach introduced by Calabrese and Paninski (2011), which consists in maximizing the log-likelihood criterion associated to the classical Gaussian mixture model (GMM)