Pattern Recognition 76 (2018) 391–403

Contents lists available at ScienceDirect

Pattern Recognition

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

Bi-weighted ensemble via HMM-based approaches for temporal data clustering

Yun Yang[a], Jianmin Jiang[b,*]

[a] National Pilot School of Software, Yunnan University, China
[b] Research Institute for Future Media Computing, College of Computer Science & Software Engineering, Shenzhen University, China

ARTICLE INFO

Article history:
Received 22 November 2016
Revised 18 October 2017
Accepted 18 November 2017
Available online 21 November 2017

Keywords:
Data clustering
Ensemble learning
Hidden Markov Model
Model selection

ABSTRACT

To improve the performance of ensemble techniques for temporal data clustering, we propose a novel bi-weighted ensemble in this paper to solve the initialization and automated model selection problems encountered by all HMM-based clustering techniques and their applications. Our proposed ensemble features in a bi-weighting scheme in the process of examining each partition and optimizing consensus function on these input partitions in accordance with their level of importance. Within our proposed scheme, the multiple partitions, generated by HMM-based K-models under different initializations, are optimally re-consolidated into a representation of bi-weighted hypergraph, and the final consensus partition is generated and optimized via the agglomerative clustering algorithm in association with a dendrogram-based similarity partitioning (DSPA). In comparison with the existing state of the arts, our proposed approach not only achieves the advantage that the number of clusters can be automatically determined, but also the superior clustering performances on a range of temporal datasets, including synthetic dataset, time series benchmark, and real-world motion trajectory datasets.

1. Introduction

As an important and unsupervised data mining technique, clustering aims at dividing an unlabeled dataset into different groups named as clusters, where all the data points grouped in the same cluster should be coherent or homogeneous [1]. While many clustering algorithms have been developed for temporal data mining tasks [2,3], they share a common trend that temporal information processing is primarily carried out via modifying the existing clustering approaches. As many such algorithms work directly with raw temporal data, these clustering algorithms are thus called proximity-based approaches, in which the major modification lies in the fact that the distance/similarity measure for static data is replaced by one designed to be more appropriate for temporal data. Other developed algorithms [4–20] intend to convert raw temporal data into either a feature vector with lower dimensions, or a number of statistical models, to which a conventional clustering algorithm is applied. These methods are often called feature-based and model-based approaches.

As indicated in the published literature [4–7,9–13], the model-based approach is widely used for temporal data clustering, in which each cluster is mathematically represented by a parametric model, such as Gaussian model [4], Hidden Markov Model (HMM) [5,6], Autoregressive Moving-Average model (ARMA) [7], Autoregressive Integrated Moving Average Model (ARIMA) [8], Mixture of Markov Chain [9,10], and fuzzy-based estimation model [11]. In such approaches, the model structure (e.g., the number of mixture models) can be determined by model selection techniques and model parameters estimated via maximum likelihood algorithms. One representative example is the well-known expectation-maximization (EM) algorithm [21].

While HMM-based clustering [5,6,12,13] has been studied for the last decade, which is regarded as an important model-based approach for temporal data analysis, most of HMM-based clustering algorithms still face the challenge that model selection and initialization sensitivity remains to be an unsolved problem. This is originally inherited from conventional partitioning and hierarchical clustering approaches. Although many model selection techniques have been reported in the literature [22–28], no one has been well accepted for general clustering. This is due to the fact that there exists no general and formal definition of what the "true clusters" are, and clustering algorithms intend to group the target dataset with different aims of cluster analysis [29]. In principle, existing approaches of determining the number of clusters can be classified into two categories: validation index determination and adaptive learning determination. In the first category, the number of clusters is externally determined by optimizing the

* Corresponding author.
E-mail addresses: yangyun@ynu.edu.cn (Y. Yang), jianmin.jiang@szu.edu.cn (J. Jiang).
URL: http://futuremedia.szu.edu.cn (J. Jiang)

https://doi.org/10.1016/j.patcog.2017.11.022
0031-3203/

pre-defined criterion, such as Akaike Information Criterion (AIC) [22], Bayesian Information Criterion (BIC) [23], minimizing description length [10] and minimization function for shape and texture clustering [24]. Recent empirical studies on model selection [30,31] reveals, however, that such method either over-estimates or under-estimates the intrinsic number of clusters in reality. In the second category, clustering algorithm itself is able to gradually update the structure of clusters during an iterative learning process, and the number of clusters can be automatically determined until a stop criterion is reached. Typical examples include DBSCAN [25], Adaptive K-means [26], fuzzy adaptive clustering [27], and adaptive fuzzy C-means clustering [28]. Although such approaches have shown promising results to an extent, most of them still suffer from the problem of initialization sensitivity, leading to higher computational cost in practical applications

Ensemble learning techniques have been widely used in clustering [5,32–36], in which the so-called clustering ensemble approaches attempt to improve the robustness of clustering by combining multiple clustering solutions into a single consolidated clustering ensemble. Such application normally achieves better results in terms of average performance among given clustering solutions, leading to a potential solution for the initialization problem. Although such techniques have been intensively researched and developed, how to harmonically combine the various clustering solutions into an optimal consensus without any prior information is still a serious challenge [37].

Built upon our previous work [15,16,38], we propose a HMM-based ensemble for temporal data clustering, in which the ensemble technique is used to tackle the initialization problem caused by HMM-based K-models during the initial clustering analysis. In comparison with the existing state of the arts, our proposed provides an optimal reconciliation for input partitions through a so-called bi-weighting scheme, where the two sets of weights for the partitions and clusters are assigned in accordance with their level of importance during the learning process throughout all iterations. By applying the DSPA (*dendrogram-based similarity partitioning* algorithm) consensus function [15], the proposed bi-weighting scheme is able to optimize the final partition with intrinsic number of clusters and hence improves the temporal data clustering performances. To this end, our contribution can be highlighted as: (i) We propose an improved HMM-based clustering ensemble and hence provide a new solution for the initialization problem and model selection problem encountered by all temporal data clustering algorithms as well as their applications; (ii) We introduce a novel optimization scheme to transform the input partitions into a single consolidated clustering solution, where level of importance is considered and adapted during the learning and transformation process by a so-called bi-weighting scheme.

The rest of paper is organized as follows. Section 2 reviews HMM-based clustering as background knowledge of designing our approach. Section 3 describes our approach, together with the detail of major techniques developed. Section 4 reports the experimental results on various temporal datasets. Section 5 discusses the issues related to our approach, and finally, the conclusions are drawn in Section 6.

2. Overview of HMM-based clustering

To pave the way for our proposed algorithm, we overview the HMM-based representation of temporal data and the existing HMM-based clustering algorithms in this section.

2.1. HMM-based representation of temporal data and K-models

HMM describes an unobservable stochastic process consisting of a sequence of states $\{q_t\}_{t=1}^{T}$ with $q_t \in \{1, 2, ..., S\}$, each of which is related to another stochastic process that generates observations of temporal data $x = \{o_t\}_{t=1}^{T}$. Initially, an observation o_1 is generated with an emission probability $b_i(o_1) = p(o_1|q_s = i)$ at the state i, which is selected according to the initial probability $\pi_i = p(i = 1)$. The next state j is determined by the state transition probability a_{ij}, and an observation o_2 is also generated based on an emission probability $b_j(o_2) = p(o_2|q_s = j)$ at the state j. The process repeats until a sequence of observations $\{o_t\}_{t=1}^{T}$ are generated. Essentially, the entire process produces a sequence of observations instead of the states, from which the name 'hidden' is drawn. The complete set of HMM parameters is described by a triplet $\lambda = \{\pi, A, B\}$, where $\pi = \{\pi_i\}_{i=1}^{S}, A = \{a_{ij}\}_{i=1,j=1}^{S,S}, B = \{b_i\}_{i=1}^{S}$. For continuously valued temporal datasets such as time series, it is normally assumed that each state generates observations according to a multivariate Gaussian distribution, due to fact of that there are efficient parameter estimators for this special case. However, such assumption does not hold true for all kinds of temporal data, and as a result, extensive research [39–42] has been carried out for non-Gaussian emission distributions. Without prior information of the target datasets, however, the choice of emission distributions is uncertain, and hence this remains to be an unsolved problem.

In our approach, we assume that the emission distribution of continuously valued temporal data is modelled as a single Gaussian distribution $b_i = \{\mu_i, \sigma_i^2\}$. Extensive experiments and analysis support that such a single Gaussian emission distribution can reduce the computational cost, and prevent the risk of over-fitting for HMM modelling. Correspondingly, a temporal dataset can be modelled as a set of K HMMs $\Lambda = \{\lambda_1, \lambda_2, ...\lambda_K\}$ with S states based on single Gaussian distributed observations, where λ_k represents a cluster of temporal data with three model parameters: (i) An initial state distribution $\pi^k = \{\pi_i^k\}_{i=1}^{S}$, (ii) A state transition probability matrix $A^k = \{a_{ij}^k\}_{i=1,j=1}^{S,S}$, (iii) S observation emissions in single Gaussian with mean: $\{\mu_1^k, \mu_2^k, \ldots \mu_S^k\}$ and variance: $\{\sigma_1^{k2}, \sigma_2^{k2}, \ldots \sigma_S^{k2}\}$.

As a general form of K-means, HMM-based K-models [6] model the entire dataset $X = \{x_n\}_{n=1}^{N}$ as a set of K HMMs $\{\lambda_k\}_{k=1}^{k}$ with S states. The parameters of K HMMs with pre-defined number of states S are initially estimated on K data points, which are randomly selected from the target dataset without any replacement. A log-likelihood of each data point under K HMMs is then calculated by the Forward and Backward algorithms [43,44], and each of them is assigned to the HMM with maximum log-likelihood. After that, the parameters of K HMMs are re-estimated on the corresponding cluster of data points by EM algorithm [21]. The entire process is repeated until the cluster memberships no longer change.

2.2. HMM-based agglomerative clustering

Originally proposed by Smyth [12], HMM-based agglomerative clustering incorporates an adaptive training process in which each data item is initially treated as a cluster represented by a singleton HMM, and N singleton HMMs $\{\lambda_N\}_{n=1}^{N}$ are trained on the entire dataset $X = \{x_n\}_{n=1}^{N}$. The closest pair of clusters, indicated by i and j, are merged as a new cluster k represented by a composite model, and the composite model is represented by the parameters of its children models $\lambda_k = \{\lambda_i, \lambda_j\}$. This process is repeated until a stop criterion such as pre-defined number of clusters is reached. During each iteration, the closest pair of cluster i and j are chosen to merged according to *Kullback–Leibler* (KL) distance measurement [45]:

$$D_{KL}(\lambda_i, \lambda_j) = \sum_x p(\lambda_i)\left[\log p(x|\lambda_i) - \log p(x|\lambda_j)\right] \quad (1)$$

In our approach, the distance between two clusters i and j is defined as a symmetric version of the KL distance, details of which