Experimental Analysis of the Behavior of Short Time Series Clustering

Marta Quemada-Lopez[1] Miho Ohsaki[1] Shigeru Katagiri[1]

Abstract: Time series clustering (TSC), which is an important technique to discover groups of time series and their representatives, has wide applications. There are several effective model-based approaches to TSC. However, since time series, especially those of biomedical observations, are often too short to be modeled appropriately, clustering such short time series remains a difficult research issue. To solve this problem, we select two major model-based TSC approaches, deterministic and probabilistic, and analyze their behaviors using simulated short time series. Experimental results suggested that the superiority of these two approaches depended on the inter-cluster distances of the clusters.

1. Introduction

Time series clustering (TSC) is an exploratory procedure to discover groups of time series and their representatives. Focusing on biomedical fields, TSC applications can analyze microarray gene expressions, mitotic cell cycles, and the symptoms of chronic diseases. TSC is difficult for such time series, because they have from several to a few dozen points, and lack information associated with this short length. Even though this problem has been specifically discussed for each application, discussion from a unified view of similar applications is also needed. This study analyzes TSC behavior for short time series using simulated data to obtain guidelines about which TSC method is effective under which conditions.

Although there are various approaches to TSC, the model-based one is common for the analysis of biomedical time series because of its ability to discover the stochastic mechanisms behind time series [1], [2]. A model-based TSC can be categorized into deterministic and probabilistic approaches; one assigns time series to clusters using the distances among the time series models [3], and the other estimates the probabilities that the models belong to clusters [4]. We experimentally examined how these two methods worked, depending on different time series lengths and different inter-cluster distances.

2. Model-based Time Series Clustering

Assuming that an underlying stochastic process generates time series, a model-based TSC organizes groups of time series whose modeled processes are similar: in other words, those that might have been generated by one particular process. For biomedical time series, the family of autoregressive (AR) models is commonly applied, including AR moving average (ARMA) and AR-integrated MA (ARIMA) [3], [4].

There are two approaches to model-based TSC, deterministic and probabilistic, and the methods based on them that we selected [3], [4] actually used the AR family. The deterministic [3] and probabilistic methods [4] were respectively formulated using ARIMA and ARMA. The ARIMA model can be replaced by the ARMA model by the transformation of a time series that eliminates trends and seasonality through differencing. The present study clarifies the effectiveness of the clustering approaches themselves, and hence the ARMA model was commonly used for both methods. The ARMA model is defined in Eq. (1), where $x[t]$ is the point at time t on time series $\mathbf{x}$, ϕ_0 is the bias term, $\phi_0, \phi_1, \phi_2, ..., \phi_p$ are the AR coefficients, $\theta_1, \theta_2, ..., \theta_q$ are the MA coefficients, and σ^2 is the variance of Gaussian noise $N(0, \sigma^2)$:

$$x[t] = \phi_0 + \sum_{i=1}^{p} \phi_i x[t-i] + \sum_{j=1}^{q} \theta_j e[t-j] + e[t] \qquad (1)$$

$$e[t] \overset{\text{iid}}{\sim} N(0, \sigma^2).$$

2.1 Deterministic Approach

The TSC method, which is based on the deterministic approach [3], assigns time series to clusters using the distances between models of time series. This hard assignment can be represented by the $N \times K$ matrix $\mathbf{A}$ shown in Eq. (2), where N is the number of time series, K is the number of clusters, a_{nk} is a component indicating whether n-th time series $\mathbf{x}_n$ belongs to k-th cluster ω_k by 1 or not by 0:

$$\mathbf{A} = [a_{nk}]_{N \times K}. \qquad (2)$$

The objective function is the sum of the within-cluster distances defined in Eq. (3) $J(\mathbf{A})$, where $\mathbf{m}_k$ is the medoid of the k-th cluster, and $d(\mathbf{x}_n, \mathbf{m}_k)$ is the distance of the n-th time series to the medoid. This is the Euclidean distance between AR-based cepstra, aka linear predictive coding (LPC) cepstra. The LPC cepstrum, which is a good representation of the amplitude spectrum envelope of a time series, is derived through ARMA modeling

[1] Doshisha University
1-3 Tataramiyakodani, Kyotanabe-shi, Kyoto 610-0321, Japan

and the conversion of the ARMA coefficients:

$$J(\mathbf{A}) = \sum_{n=1}^{N} \sum_{k=1}^{K} a_{nk} d(\mathbf{x}_n, \mathbf{m}_k). \qquad (3)$$

The search for the best assignment, i.e., the minimization of the distance-based objective function, is done by a common k-medoids algorithm called partitioning around medoids.

2.2 Probabilistic Approach

The TSC method, based on the probabilistic approach [4], estimates the probabilities of the models of time series that belong to each cluster. This soft assignment can be represented by the $N \times K$ matrix shown in Eq. (4), where $\mathbf{\Phi}$ denotes a set of a bias term and ARMA coefficients: $\{\phi_0, \phi_1, \phi_2, \cdots, \phi_p, \theta_1, \theta_2, \cdots, \theta_q\}$, and hence, $\mathbf{\Phi}_k$ is a model that represents the k-th cluster. Component $\Pr(\mathbf{x}_n \mid \omega_k, \mathbf{\Phi}_k)\ \Pr(\omega_k)$ denotes the probability that the k-th cluster occurs, whose model is described by $\mathbf{\Phi}_k$, and the n-th time series originates from this model:

$$\mathbf{P} = [\Pr(\mathbf{x}_n \mid \omega_k, \mathbf{\Phi}_k)\Pr(\omega_k)]_{N\times K}. \qquad (4)$$

The objective function is the log likelihood defined in Eq. (5) $J(\mathbf{P})$, where $\mathbf{\Phi}_n$ is a set of the bias term and ARMA coefficients of the model of the n-th time series. Since the distributions of all the probabilistic variables are assumed to be Gaussian, the maximization of this likelihood-based objective function is equivalent to the estimation of a Gaussian mixture:

$$J(\mathbf{P}) = \sum_{n=1}^{N} \left(\ln \sum_{k=1}^{K} \Pr(\mathbf{x}_n \mid \omega_k, \mathbf{\Phi}_k)\Pr(\omega_k) \right). \qquad (5)$$

The EM algorithm is employed to optimize the parameters of modeling and clustering, and in this process, there are two possible implementations of the calculation of $\mathbf{\Phi}_k$. Regarding time series $\mathbf{x}_n$ that is assigned to the k-th cluster, one is updating $\mathbf{\Phi}_k$ based on the model parameters of $\mathbf{x}_n$, and the other is retraining $\mathbf{\Phi}_k$ by inputting $\mathbf{x}_n$. For the consistency in modeling procedures between the deterministic and probabilistic methods, we selected the former implementation, which was previously used [3].

3. Comparative Experiments

3.1 Purpose and Conditions

We experimentally evaluated and compared the two approaches using simulated data. The ARMA model can be theoretically converted into the AR model, and this conversion was done in previous experiments [3], [4]. Following this manner, second order AR models were used for generating and modeling the simulated data.

To examine the effects of the distance between clusters and the length of the time series, these factors were controlled as below. We respectively generated three datasets, 1, 2, and 3, in which the inter-cluster distance was large, medium, and small. For each dataset, the following process to generate time series was iterated by changing its length from 5 to 100 time points. As true representatives of the clusters, we used the coefficients of eight AR models plotted in Fig. 1 whose values were based on previous work [2]. Four of the eight AR models corresponded to four clusters in each dataset, and each model generated 25 time series.

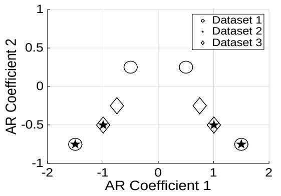

Fig. 1 AR coefficients used for each dataset

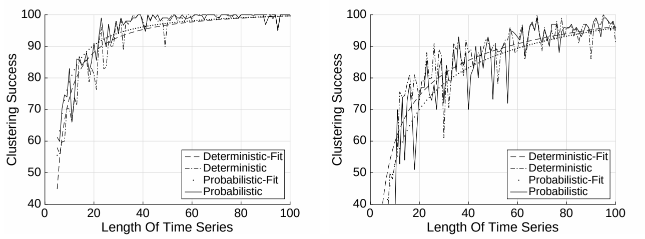

Fig. 2 Experimental results for Datasets 1 (left) and 2 (right). Larger values denote better performance.

For revealing general trends, we repeated the clustering under 100 different initializations and averaged the results. For clarifying the curve over the time series length, the averaged results were fitted by a function. Both power and exponential functions were tried, and the former was selected for fitting due to its better performance.

3.2 Results and Discussion

Fig. 2 shows the clustering success rates that were experimentally obtained. As shown for Dataset 1, the probabilistic method worked better when the clusters were far away from each other (see the dotted line). For Dataset 2, the deterministic method performed better when the clusters were at a medium distance from each other (see the dashed line). No difference appeared between the two approaches for Dataset 3 when the clusters were too close (this result was omitted). The absolute performances of both methods were low when the time series were very short.

4. Conclusions

To analyze how model-based TSC behaves for short time series, we experimentally examined its two major approaches. The probabilistic approach was superior for clusters with large inter-cluster distances, while the deterministic approach was superior for comparatively small distances. However, their performances need improvement because they were considerably low for very short time series. Future work will propose a new method specific to such time series, based on the knowledge of the two approaches gleaned from this study.

References

[1] J. Ernst and Z. Bar-Joseph, SSTEM: a Tool for the Analysis of Short Time Series Gene Expression Data, BMC Bioinformatics, vol. 7(1), pp. 1-11 (2016).

[2] E. D. Foster: State Space Time Series Clustering Using Discrepancies Based on Kullback-Leibler Information and the Mahalanobis Distance, PhD Thesis, University of Iowa (2012).

[3] K. Kalpakis et al., Distance Measures for Effective Clustering of ARIMA Time-Series, IEEE Int'l Conf. on Data Mining ICDM-2001, pp.273-280 (2001).

[4] Y. Xiong and D.-Y. Yeung, Mixtures of ARMA Models for Model-Based Time Series Clustering, IEEE Int'l Conf. on Data Mining ICDM-2002, pp. 717-720 (2002).

This work was supported by a MEXT-supported program for the Strategic Research Foundation at Private Universities 2014-2018.