Comparison of Short Time Series Clustering Methods

Marta Quemada López † **Miho Ohsaki** † **Shigeru Katagiri** †
† **Graduate School of Science and Engineering, Doshisha University**

1. Introduction

Time series clustering (TSC) is an important tool for knowledge discovery and data mining in many fields like socioeconomics, environmental, and biomedical sciences. However, time series sometimes present lack of information because of short length. This aspect has not been discussed much yet. Therefore, we compare two major TSC approaches using simulated short time series and clarify which one of them works better.

2. Compared Methods

When time series can de represented by a stochastic process, it is reasonable to select a TSC method, which assumes the following: time series belong to the same group when the underlying mechanisms that produced them are similar. Such TSC methods are categorized into two approaches. One assigns time series to clusters in a deterministic way, and the other does it in a probabilistic way. We attempt to compare these two approaches using the methods below.

A method based on the deterministic approach was proposed by Kalpakis et al. (2001) [1]. It adopts autoregressive moving average (ARMA) model to estimate the underlying stochastic mechanism. It organizes clusters using the distance of the cepstral coefficients and the partitioning around medoids algorithm.

The second method, based on a probabilistic approach, was proposed by Xiong and Yeung (2002) [2]. It adopts the same stochastic model as Kalpakis, and estimates the mixing coefficients of the Gaussian mixture distribution of ARMA models with the expectation-maximization algorithm.

3. Experiments

In order to compare the two methods four different artificial datasets were generated. For each dataset we simulated four clusters, and generated time series corresponding to each cluster. Each cluster was defined by a second order autoregressive model and contained 25 time series with length ranged from 10 to 20 time points. Both methods were applied to these datasets 10 times, and the results were averaged over these ten trials.

Dataset		1	2	3	4
Deterministic	A (%)	73.4	74.0	71.5	61.5
	B	0.385	0.613	0.182	0.158
Probabilistic	A (%)	66.5	64.3	59.2	54.3
	B	0.491	0.482	0.450	0.290

Table 1. Experimental Results
A: Percent of correctly clustered – the higher the better
B: Estimation error of cluster representatives – the lower the better

The experimental results in Table 1 show that the deterministic method worked better than the probabilistic one for short time series. It was suggested that the smaller number of cluster parameters of the deterministic method results in the better performance.

4. Summary

We examined two representative methods for time series clustering using short time series. As a result, the deterministic method outperformed the probabilistic one. Our future work will focus on the proposal of a new deterministic method for higher performance.

Acknowledgements

This work was supported in part by MEXT-Supported Program “Driver-in-the-Loop”.

Bibliography

[1] 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).

[2] 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).