Symbolic time series notation for time series clustering

Jerzy Korzeniewski

University of Lodz, Poland.
Phone +4842-6355182, e-mail jurkor@wp.pl

1. Introduction – This presentation contains a proposal of a novel algorithm which may be useful in time series data analysis. The aim of the algorithm is to reduce the size of all time series in a given data base in such a way that their cluster structure (if there is one) would still be retrievable. The first step of the algorithm is the one which is almost obligatory in all approaches to this task i.e. finding a much smaller number of segments which represent the time series. To this end we use the PAA (*piecewise aggregate approximation*) technique not specifying the exact number of segments but allowing it to vary from 10 to 30. For each segment we find the arithmetic mean of all observations comprised by the segment. The second step of the algorithm consists in identifying more important time series segments by means of investigating distance based correlation between them. Investigating correlation might make sense as the number of variables (segments) is not very big any more. Distance based correlation is very useful in the context of clustering, because, if there is a cluster structure then "jumps" of objects from one cluster to another should be reflected on both sets of variables describing objects i.e. the correlation between these two sets of variables should be positive. Once the importance of all segments has been established we save (this constitutes the algorithm's third step) the whole time series in the form of binary variables assigned to each segment (not the means) which allows to further reduce the used memory size. The binary variables can be of two sorts. The first sort is defined by the monotonicity of segments i.e. when there was a rise of the values representing two consecutive segments the binary variable assumes value 1, if there was a decline the variable assumes value 0. The second sort is defined by the positioning of the value representing a given segment with respect to the overall mean. If the segment value is above the mean the binary variable assumes value 1, otherwise it assumes value 0. In the fourth step of the algorithm we select only the most important half of all the binary variables to finally represent the whole time series.

2. Experimental - The algorithm was evaluated on some time series data available from *the UCR Time Series Classification Archive* (www.cs.ucr.edu/~eamonn/time_series_data). We examined 30 data sets. We compared the clusterings based on a couple of variable sets with the proper clustering by means of the adjusted Rand index.

3. Results and Discussion - . In almost all cases the results were very good. There was no significant difference between the clustering based on the selected half of the binary variables and the clustering based on all binary variables or even on the continuous variables made up by the means of all segments. In addition the algorithm provides a reasonable choice of the number of segments to represent the time series. This number is chosen by maximizing the mean distance based correlation coefficient.

4. Conclusions – The algorithm seems to work well in the task of time series clustering in spite of the weakening of the measurement scale. This result coincides with similar conclusions of other researchers. The algorithm's advantageous feature is that it can be applied to time series measured on any scale.

5. References

[1] E. Keogh, K. Chakrabarti., M. Pazzani, et al., *Dimensionality reduction for fast similarity search in large time series databases*, Knowledge and Information Systems 3 (3), (2000) p. 263–286.

[2] J. Korzeniewski, *Metody selekcji zmiennych w analizie skupień. Nowe procedury.* Wydawnictwo Uniwersytetu Łódzkiego, 2012.

[3] Ratanamahatana C., Keogh E., Bagnall A., et al., *A novel bit level time series representation with implications for similarity search and clustering*, in: PAKDD, Hanoi, Springer, (2005), p. 771–777.