Master thesis/research project topic

Clustering time series data by compression

Context

Clustering is the well-established problem of grouping elements according to their similarities. It has been very successful in grouping information, mostly in areas where a distance or similarity between elements is well-defined. However, for certain data types, such as time series, this definition is lacking. In these cases a common way of measuring similarity between elements is by associating a model to each element and then measure the difference between the models. For probabilistic and information theoretic models this distance is well-defined and given by the Kullback-Leibler divergence.

Algorithmic information theory tells us that the more structure the data has, the more it can be compressed. This means that structured data is made of small 'building blocks' (or regularities) that repeatedly occur. Compression-based clustering has been very successful at grouping languages, literature, animal species, music, etc. and it has been continually applied to new types of data.

Goal

The goal of this work is to develop a new clustering technique for time series based on the Minimum Description Length (MDL), an information theoretic principle that permits to trade-off model complexity with goodness of fit. A literature study of compression-based clustering should be made to not only understand the landscape of the field but also derive insights for the new method. In the end, the new clustering algorithm should be validated with synthetic datasets and real world datasets, such as heart rate for social anxiety or EEG for sleep patterns.

Research questions

The main research question could be, for example:

- Which model is most appropriate to represent time series in the context of clustering?

Sub-questions could include:

- What represents similarity between the time series models?
- How can we expand on the previous research?
- How to represent the model in MDL?

Realisation

- Literature study
- Define possible models
- Implement the method (theoretically and practically)
- Execute the new method in synthetic and real world data
- Compare with existing algorithms
- Evaluate the results, derive recommendations.
- Write report / thesis.

Student profile

Good understanding of data mining problems and algorithms; theoretically oriented; experience/interest in information theory and data structures; preferable programming experience in Python or R.

Supervisors

Hugo Manuel Proença (h.manuel.proenca@liacs.leidenuniv.nl)
Matthijs van Leeuwen (m.van.leeuwen@liacs.leidenuniv.nl)

Relevant literature:

[1] Classic clustering by compression:
Cilibrasi, Rudi, and Paul MB Vitányi. "Clustering by compression." *IEEE Transactions on Information theory* 51.4 (2005): 1523-1545.
http://ieeexplore.ieee.org/abstract/document/1412045/

[2] Summarizing transaction datasets with MDL (Krimp):
Vreeken, Jilles, Matthijs Van Leeuwen, and Arno Siebes. "Krimp: mining itemsets that compress." *Data Mining and Knowledge Discovery* 23.1 (2011): 169-214.
https://link.springer.com/article/10.1007%2Fs10618-010-0202-x?LI=true

[3] Applying MDL clustering to Seismographs:
Bertens, Roel, and Arno Siebes. "Characterising seismic data." *Proceedings of the 2014 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, 2014.
http://epubs.siam.org/doi/abs/10.1137/1.9781611973440.101

[4] MDL modelling of multivariate time series:
Bertens, Roel, Jilles Vreeken, and Arno Siebes. "Keeping it short and simple: Summarising complex event sequences with multivariate patterns." *arXiv preprint arXiv:1512.07056* (2015).
https://arxiv.org/abs/1512.07056

[5] Clustering multivariate time series with Hidden Markov Models:
Ghassempour, Shima, Federico Girosi, and Anthony Maeder. "Clustering multivariate time series using hidden Markov models." *International journal of environmental research and public health* 11.3 (2014): 2741-2763.
http://www.mdpi.com/1660-4601/11/3/2741/htm

[6] MDL for time series:
Ugo Vespier, Arno J. Knobbe, Siegfried Nijssen, Joaquin Vanschoren. “MDL-Based Analysis of Time Series at Multiple Time-Scales”. ECML/PKDD (2) 2012: 371-386
https://link.springer.com/chapter/10.1007%2F978-3-642-33486-3_24