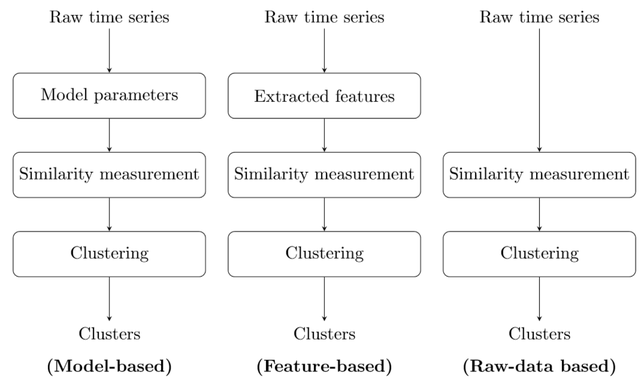
Whole-series Time-series Clustering Techniques

Yohann Jacob Sandvik
Institute of Electronic Systems - NTNU

Overview of Whole-series Time-series Clustering Techniques

- There are three types of time-series clustering, *whole-series time-series clustering*, *subsequence time-series clustering* and *time-point time-series clustering*.
- In this review we will only consider work using whole-series time-series clustering.
- Whole series time-series clustering can broadly be divided into three main approaches. The raw-data based approach, the feature-based approach and the model based approach.
- When clustering raw time series the majority of the work goes into selection of similarity metric and clustering algorithm, and one clusters the time series with regard to similarity in time or similarity in shape.
- In the feature-based approach one also clusters time series with regard to similarity in time, and shape, but the work is somewhat shifted away from choice of similarity metric and over to choice of representation.
- In the model-based approach the goal is most often to cluster time series with regard to the underlying data generating process. The underlying assumption being that two time series that appear different might still have been generated by the same process.



ARMA models

$$X_t - \phi_1 X_{t-1} - ... - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + ... + \theta_q Z_{t-q} \quad (1)$$

Hidden Markov Models

Transition probabilities.

$$\begin{aligned} p_{ij} &= P(X_n = i | X_{n-1} = i_{n-1}, ..., X_0 = i_0) \\ &= P(X_n = i | X_{n-1} = i_{n-1}) \end{aligned} \quad (2)$$

Hidden states with emmission probabilities.

$$P(S = s | X = j) \quad (3)$$

PCA and ICA

PCA and ICA are both ways of projecting the input matrix onto a reduced feature space. In PCA one finds the principal components that are the eigenvectors of the covariance matrix if the input vectors. In ICA one assumes that the matrix of observed variables is a linear combination of mutually independent variables. Hence, one tries to reconstruct a set of variables that are as "mutually independent" as possible.

Feature-based representations

Feature extraction method	Articles
Basic signal statistics	[57–60]
PCA or ICA	[58, 59, 61–65]
Time-frequency decomposition	[66–73]
Matrix, or tensor decomposition	[61, 74–77]
Symbolic aggregate approximation	[66, 78, 79]
Permutation based coding	[80]
Spline functions	[81]
Topological feature extraction	[82]
Autoencoder	[60]

Model-based representations

Time series model	Articles
ARMA model.	[54, 66, 83–89]
HMM.	[90–92]
State space model.	[62, 81, 93]
Varience ratio statistics.	[94]
Copula based model.	[95]
Network model.	[96, 97]

K-means

- Family of hard clustering algorithms.
- Number of clusters -> predetermined.
- Iterative algorithm.
- Not deterministic.

Fuzzy C-means

- Family of soft clustering algorithms.
- Number of clusters -> predetermined.
- Iterative algorithm.
- Not deterministic.

Hierarchical Clustering

- Family of hard clustering algorithms.
- Number of cluster -|> predetermined.
- Builds hierarchies of clusters.
- Similarity between clusters is defined by distance metric, and *linkage*.
- Deterministic.

Expectation Maximization

- Iterative algorithm used for estimating model parameters satisfying the ML criterion.
- Is used for clustering by letting each cluster being a generative process.
- Time and storage complexity increase exponentially with time-series length.

Discussion

- PCA and ICA are considered to be the most viable options for feature-based representations. Because they build on strong theoretical ground, without being overly complex methods they are fairly interpretable.
- Of the model-based representation methods ARMA models were found to be a frequent model used. Since they are also frequently used for feature extraction in wind turbine monitoring they should be considered for a master thesis.
- The HMM is a model-based approach that has showed great promise in terms of representing time series, and has even be used to cluster the time series [91] in a simple manner.
- SOMs have been shown to work well on financial time series by [57, 93], so they could be appropriate for the time series produced by wind turbines. However, ANNs in general require careful design and and long training, so it might take a lot of time away that could be used for testing other approaches. So SOM should not be considered as a primary model-based clustering approach for wind turbine time series.

Clustering Algorithms

Clustering Algorithm	Articles
Hierarchical clustering	[58, 60, 64–66, 69, 71, 73, 74, 80, 81, 88, 94]
K-means family	[59–61, 68, 69, 79, 81, 82, 84, 92, 94]
Fuzzy C-means family	[67, 75, 83–87, 95]
Expectation-maximization	[54, 60, 81, 90]
SOM	[57, 69, 93]
Spectral clustering	[60, 72, 84]
BIRCH	[60, 76]
Custom algorithms	[63, 77, 78, 89]