Clustering of Time Series Streams using the Centroid Decomposition

MSc Thesis:

Work overview:

Stream clustering is a fundamental problem in time series field. Unlike batch-mode clustering, there are two challenges in stream clustering: (i) Given that time series are continuously changing, how to incrementally update the clustering results? (ii) Given that clusters continuously evolve with the evolution of data, how to capture the cluster evolution activities?

Figure 1 shows an example of time series clustering. In this figure, three different time series from the PAMAP dataset [3] are represented. The three time series are retrieved from IMU sensors placed respectively on the hand, the chest and the shoe of a subject while doing a physical activity. The application of the clustering on these time series returns three different clusters that are recognized in this snippet as: Running, Cycling and other activities.

The aim of this thesis is to apply the Centroid Decomposition technique (CD) [1], [2] to perform evolving clustering in time series streams. The produced clustering should take into account the correlation across the input time series. Formally, The Centroid Decomposition technique decomposes an \( n \times m \) matrix \( X = [X_1 \ldots X_m] \) into an \( n \times m \) loading matrix \( L = [L_1 \ldots L_m] \).
and an $m \times m$ relevance matrix $R = [R_1 | \ldots | R_m]$ as follows:

\[
CD(X) = L, R \\
s.t. \quad X = L \times R^T \\
\quad = \sum_{i=1}^{m} L_i \times R_i^T
\]

As an outcome, you will implement a graphical application that visualizes the clustering process of the Centroid Decomposition technique and its result. The tool should also graphically illustrate the steps of the computation of CD technique.

**Work tasks:**

1. Familiarize yourself with Centroid Decomposition (CD) algorithm.
2. Implement an incremental version of CD algorithm for the clustering of time series streams.
3. Embed the autocorrelation detection into the CD algorithm.
4. Implement a graphical tool to visualize the result of CD based clustering.
5. Write a thesis that describes the algorithm and the tool.
6. Presentation of 20 minutes.

**Literature:**

5. Gong, S., Zhang, Y., and Yu, G. *Clustering Stream Data by Exploring the Evolution of Density Mountain*, in VLDB, 2018
Starting date of thesis: TBD
Ending date of thesis: TBD

University of Fribourg
Department of Informatics
Exascale Infolab

Mourad Khayati
Senior researcher