UNI FR
UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

Department of Informatics

University of Fribourg
Department of Informatics
Bd. de Pérolles 90
CH-1700 Fribourg
Phone. +41 26 300 84 65
Fax +41 26 300 97 26
http://exascale.info/

University of Fribourg, Dept. of Informatics, CH-1700 Fribourg

Student name
City

Switzerland

Mourad Khayati
Senior researcher
Phone +41 26 300 84 60
Fax +41 26 300 97 26
mkhayati@exascale.info

Fribourg, 6. Juni 2018

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.

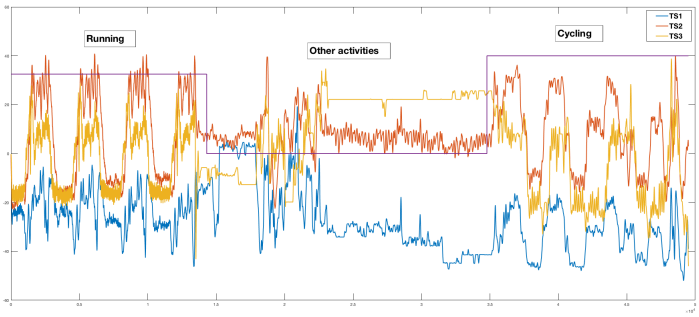

Figure 1: Example of CD based clustering of time series

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 $\mathbf{X} = [X_1|\ldots|X_m]$ into an $n \times m$ *loading* matrix $\mathbf{L} = [L_1|\ldots|L_m]$

and an $m \times m$ *relevance* matrix $\mathbf{R} = [R_1|\ldots|R_m]$ as follows:

$$\begin{aligned} CD(\mathbf{X}) &= \mathbf{L}, \mathbf{R} \\ s.t. \quad \mathbf{X} &= \mathbf{L} \times \mathbf{R}^T \\ &= \sum_{i=1}^{m} L_i \times R_i^T \end{aligned}$$

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:

1. Khayati, M., Böhlen, M.H., and Gamper, J. *Memory-efficient Centroid Decomposition for Long Time Series*, in ICDE, 2014.
2. Chu, M.T., and Funderlic, R.E. *The Centroid Decomposition: Relationships Between Discrete Variational Decompositions and SVDs*, in SIAM J. Matrix Analysis and Applications, 2002
3. http://archive.ics.uci.edu/ml/datasets/pamap2+physical+activity+monitoring
4. Nguyen, T., Gsponer, S., and Ifrim, G. *Time Series Classification by Sequence Learning in All-Subsequence Space*, in ICDE, 2017
5. Gong, S., Zhang, Y., and Yu, G. *Clustering Stream Data by Exploring the Evolution of Density Mountain*, in VLDB, 2018
6. Dong, B., et al *OnlineCM: Real-time Consensus Classification with Missing Values*, in SIAM, 2015

UNI FR
UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

Starting date of thesis: TBD

Ending date of thesis: TBD

University of Fribourg
Department of Informatics
Exascale Infolab

Mourad Khayati
Senior researcher