Ideas for the Linac3 Source ML Analysis

Max Mihailescu

This document provides a collection of ideas and topics for the analysis of the Linac3 Ion Source from a ML viewpoint. I will try to collect descriptions and links to interesting papers and summarize the results I had when trying out some of them.

Clustering

Given a collection of data points, clustering is to group together points that are similar under some kind of similarity metric. Usually, this is an unsupervised technique, meaning that no reference labels are known. There exist a variety of different algorithms, and each algorithm can produce very different results on the same set of data. Therefore it is crucial to somehow evaluate the results.

For Linac3 we performed a Clustering Analysis with the goal, to see, if certain settings of the source would lead to a stable beam current. For the resulting report please contact Detlef Küchler (CERN BE-ABP-HSL).

The clustering algorithm we used is called *Optigrid* and is described in the paper "Optimal Grid-Clustering: Towards Breaking the Curse of Dimensionality in High-Dimensional Clustering" by Alexander Hinneburg and Daniel A. Keim. (Hinneburg and Keim 1999)

Matrix Profile

The Matrix Profile is a tool for efficient motif discovery in time series, i.e. for discovery of repeated or "conserved" patterns. It can also be used to find time series discords, i.e. anomalies. A large number of papers was published by the same working group, you can find the projects website here (Keogh, n.d.). For a good introduction with example applications read the first paper (Yeh et al. 2016).

What is the Matrix Profile?

First, let's understand what a subsequence is. Consider a time series $T = T_1, \ldots, T_n$ of length n, and fix a number $m < n$ (typically $m \ll n$). Now, a subsequence of length m starting at i is the continuous block $T_i, \ldots T_{i+m}$. By sliding a window of size m over T we can get all $n - m + 1$ subsequences of length m.

Now, we can define the distance between two subsequences. The most obvious choice is the euclidean distance $dist(Q, T) = \sqrt{(Q_1 - T_1)^2 + \cdots + (Q_m - T_m)^2}$. However, for the matrix profile we work with the so called *z-normalized* euclidean distance. It is the euclidean distance

of the z-normalized subsequences, i.e. we subtract the mean of the subsequence and divide the difference by its standard deviation, and only then take the elidean distance. By this we rescale all subsequences to make them more compareable. Note that under some circumstances this might be undesirable, for example when search for motifs in the call of a bird where the pitch could be relevant.

Discretization

Change Point Detection

Suffix arrays

References

Hinneburg, A., and D. A. Keim. 1999. "Optimal Grid-Clustering: Towards Breaking the Curse of Dimensionality in High-Dimensional Clustering." *Proceedings of the 25th International Conference on Very Large Databases*, 506–17. https://kops.uni-konstanz.de/bitstream/handle/123456789/5790/vldb99.pdf.

Keogh, Eamonn. n.d. "The Ucr Matrix Profile Page." https://www.cs.ucr.edu/~eamonn/MatrixProfile.html.

Yeh, Chin-Chia Michael, Yan Zhu, Liudmila Ulanova, Nurjahan Begum, Yifei Ding, Hoang Anh Dau, Diego Furtado Silva, Abdullah Mueen, and Eamonn Keogh. 2016. "Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View That Includes Motifs, Discords and Shapelets." *IEEE ICDM 2016*. https://www.cs.ucr.edu/~eamonn/PID4481997_extend_Matrix%20Profile_I.pdf.