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 [5].

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 [6]. For a good introduction with example applications read the first paper [14].

What is the Matrix Profile?

First, let's understand what a subsequence is. Consider a time series $T = T_1 \dots T_n, \quad T_i \in \mathbb{R}$ 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 $T_i \dots T_{i+m}$. 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 the subsequence' standard deviation, and only then take the euclidean 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, and not only the shape of a signal.

The Matrix Profile P is a meta time series, that for every subsequence stores the smallest distance to some other subsequence. Furthermore, the Matrix Profile Indices I is another meta time series, that for every subsequence stores the starting index of the subsequence, that has the smallest distance to it (its nearest neighbor). If you imagine the subsequences being points in a m dimensional Space, then the matrix profile index of a subsequence is the starting index of the closest neighbor and the matrix profile value is the distance to the closest neighbor.

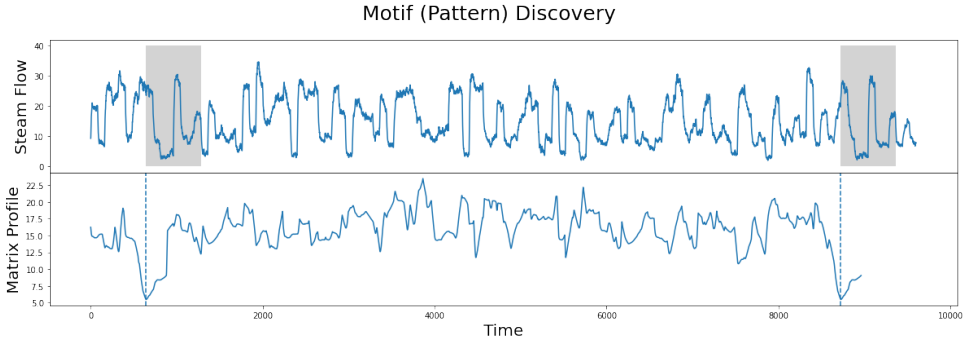

Figure 1: Example of the matrix profile on the steamgen dataset.

So, what does this tell us? In the example above you can see a visualization of the Matrix Profile for the Steamgen Dataset. In the upper plot you see the data set, and in the lower plot you see a plot of the Matrix Profile. The two dashed lines represent the lowest values in the Matrix Profile. This means the two subsequences that start at each of these lines have a very small (the smallest) distance from each other, hence they are very similar. This means, that by visually inspecting the Matrix Profile we can immediately see the most repeated pattern. There are various other things to discover and better overview can be found on the UCR page [6]. There exist also generalizations to more dimensions, see [13].

There exists a very good Python library called *stumpy* for computing the Matrix Profile that also has very good support by the author. The Github page can be found here [7].

How could the MP beused?

As described above, the MP profile is a tool that can help to discover repeating or anomalous patterns in time series data. Hence we can aim to apply it to any of the many time series produced by the source.

1. Patterns in the BCT currents for prediction: One could try to discover repeating patterns in the BCT currents and see if they can be used to predict the future development. For example if a pattern indicates a degradation of the current in the near future, it could be used to alert the operators in time. There exists also a real time version of the MP, where it gets updated with every arriving data point. For this the SDTS algorithm [12] built on top of the MP could be interesting.
2. Pattern in the BCT current for analysis: Likewise, one could try to link patterns in the BCT with patterns/actions of other parameters. For example, often times a slow increase of the HT current leads to a slow degradation of the BCT current.
3. Motifs of different parameter combinations shifted in time: When computing the multidimensional matrix profile, to see if there are motifs in more than on dimension, one could shift one of the time series in time, to see for example how a change of the gas voltage affects the current in one hour.
4. Meta time series: Instead of looking at the original time series, one could try to look for motifs or discords in a rolling window time series. For example, one could calculate the standard deviation in one hour windows over the BCT25 current and look for repeated patterns there, to maybe find motifs that will indicate a future unstable period.

Difficulties

The MP is built under the assumption, that repeated patterns or motifs are an effect of a regular event in the generating process. One example from the Papers is Seismology. There the time series is the recording of a Seismograph, which can have very long periods of "random" data, where nothing happens. However, an earthquake would show up with a very distinctive shape.

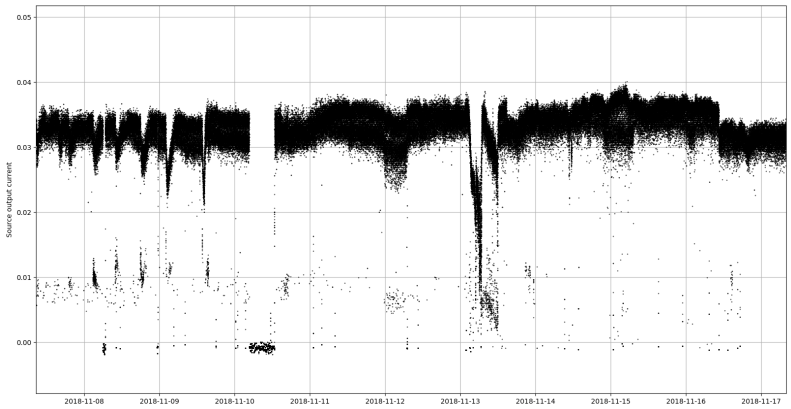

Figure 2: Example of the BCT25 current

In our case however, the BCT signals are mostly flat with oscillations (see the two images above), hence motif discovery with z-normalized subsequences is very insufficient. If you have two flat signals with a lot of added noise, their z-normalized euclidean distance will be very large, even if one might expect it to be small because visually they are very similar. There are some ideas to mitigate the problem [8], but I didn't get any satisfying results.

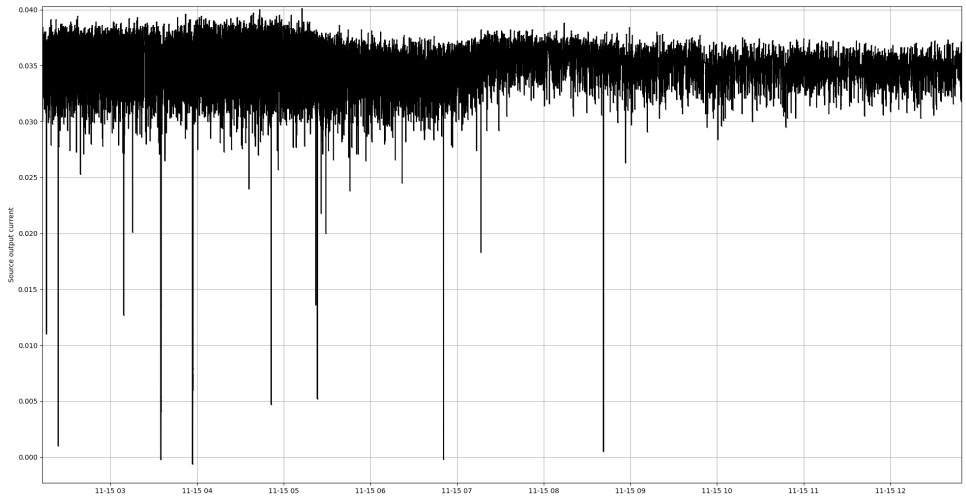

Figure 3: Example of the BCT25 current, smaller window

Results

Instead of trying to use a matrix profile with removed noise I flattened the signal over some minutes and found several links of parameter changes to BCT current. The results where achieved using the MSTOMP algorithm [13], a multi dimensional generalisation of the matrix profile calculation.

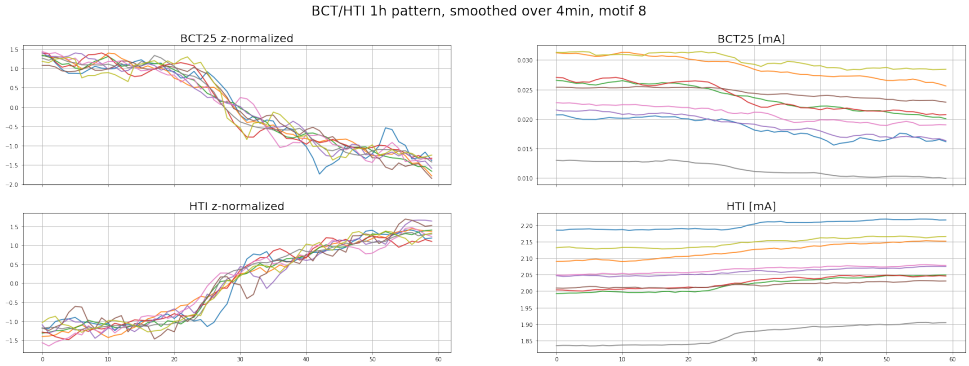

Figure 4: Rise of HTI (bottom) and degradation of BCT25 current (top).

A rising HTI often coincides with a degrading BCT25 current, so this seems to confirm this theory. However it is not a proof that an opposite effect (e.g. rising HTI and rising BCT current) doesn't also exist. But we didn't observe it in the time frame we considered (August and September 2016).

The reverse can also be seen. Drops of the HTI correlate with jumps in the BCT25 current.

One possible explanation we explored is that the behavior in the second image can often time be achieved by increasing the Oven Power. This can be seen in this image.

However, it appears that there are cases where the oven remains unchanged. Further invastigation is necessary, but it could be that the power of Oven 2 was increased, as this is only the data for Oven 1.

We could also see that an increase of the oven power is often accompanied with a decrease of the gas voltage.

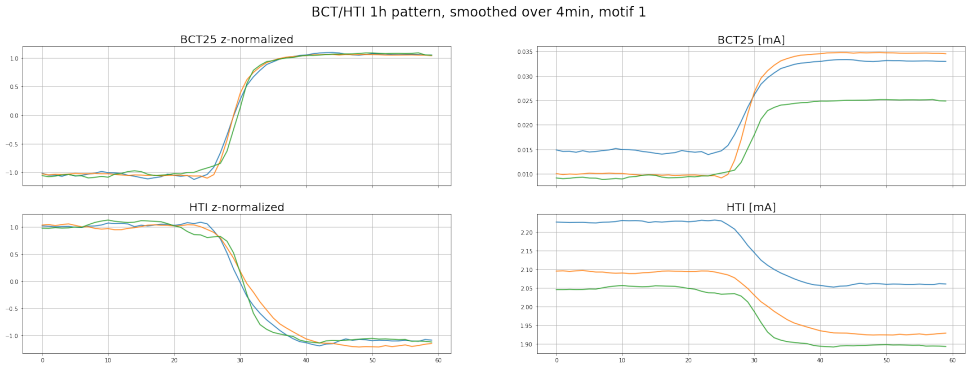

Figure 5: Drops of the HTI and jumps in the BCT25 current.

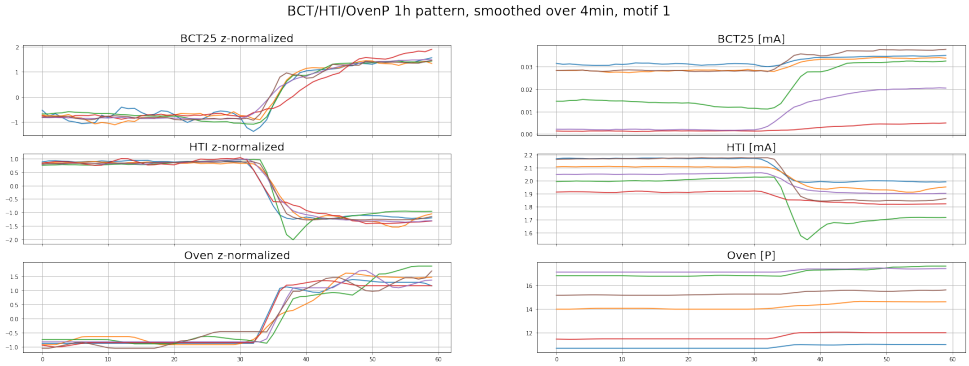

Figure 6: A common pattern when including BCT25 current, HTI and Oven1 power.

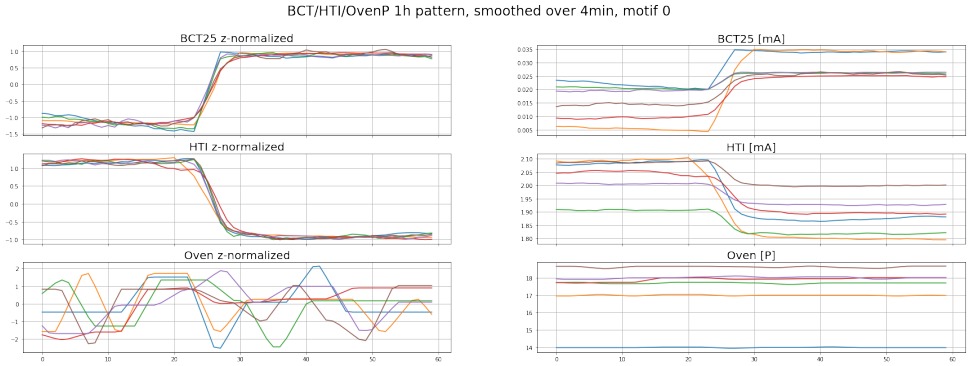

Figure 7: Jumps in BCT25 current and HTI without changes of the Oven power.

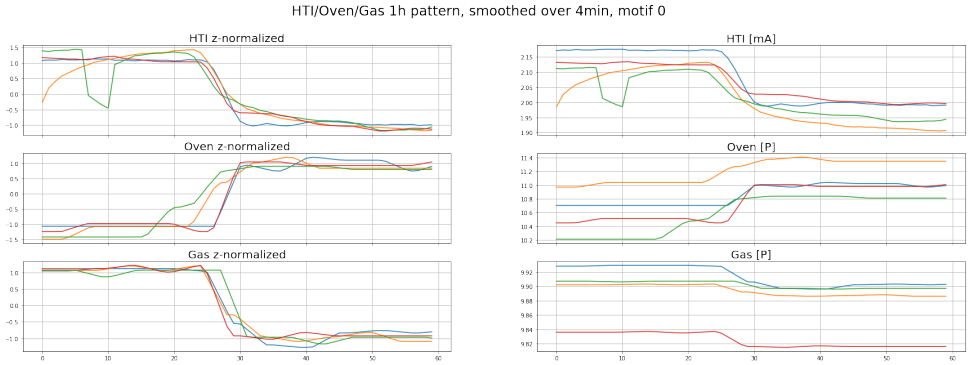

Figure 8: Correlation of gas voltage decreases with oven power increases.

We didn't see any meaningful motifs when jointly looking at the Bias Disc Voltage and the BCT25 current.

Discretization

One problem of using the data from CALS/NXCALS is that for all setting only acquisitions are logged. This means, that we typically don't know the exact values a setting was changed to, see the figure below for example. It shows one day (05.11.2018) of Oven1 power acquisition from NXCALS, where only one data point every five minutes is logged. From the logbook we can learn that at 14:00 the Oven1 Power was set to 12.0W, however on the plot we can see some oscillations (The times in the plot are UTC, so you have to count +2h).

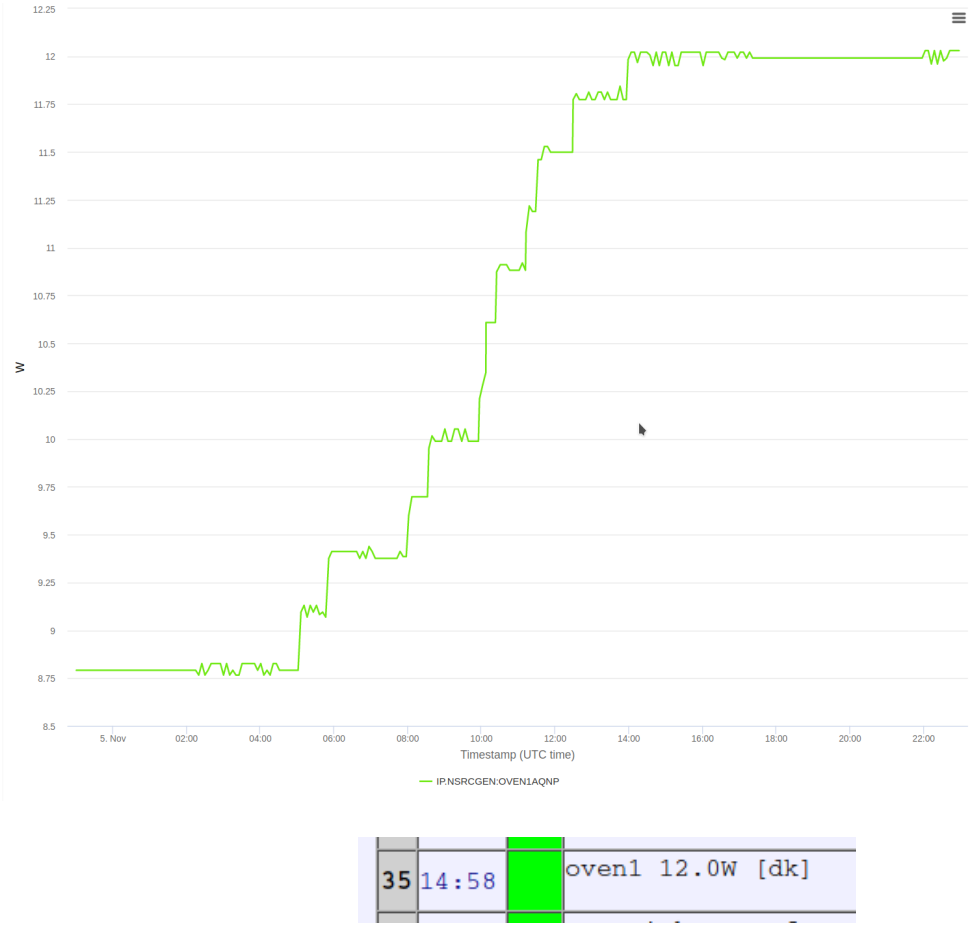

Figure 9: Top: Screenshot of Oven1 Acquisition, Bottom: The real value that was set at 14:00.

The same occurs with other settings, and raises the problem that we cannot directly say when a change of a certain setting happened. So I tried to discretize the raw acquisition values and get back the true setting where possible. The main assumption I had to make is that a setting remained constant over time, unless somebody changes it. This appears reasonable in most cases, but in some extreme cases some information might be lost (see below).

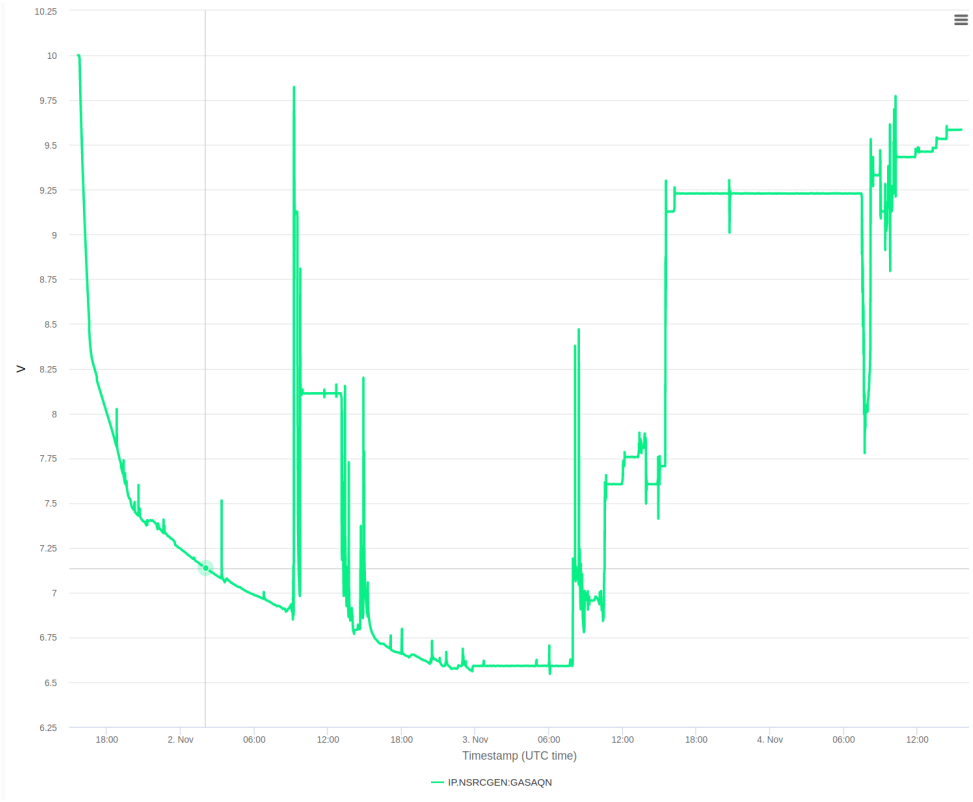

Figure 10: GASAQN voltage on the second and third November 2018. This was during an oven refill where the gas pressure changes not as a step function.

Under this assumption we can model a setting as a step function [11] with added noise, and the problem is to find the step function. I will call the step function *discretization*, because we separate our time series into discrete states of a fixed setting.

There are several techniques that could be used to solve such a problem, and we will discuss some of them below in more detail for a different use case. For this use case I combined a simple rolling window approach with a decision tree regressor. Some results can be seen below.

As one can see, the discrete approximation, our attempt at finding the true step function, follows the acquisition signal very closely and most changes are modeled correctly. This can also be seen when comparing the results with entries in the elogbook (especially for the oven, since here most changes are noted in the logbook). As can be seen in the figures, during some periods no discrete approximation is plotted. This is the case when the source was off (BCT05 current 0A), because the method does not work well when there are sections with a non-step function like signal as during an oven restart, so I cut them out.

Explanation

As described above, the process involves using a Decision Tree Regressor. A decision tree partitions the input data by sequentially applying if-then-else rules. It can be thought of as a directed graph, were every node is one of these rules. End nodes, so called leaves, that return the class the input is belonging to. Training a decision tree means finding an optimal set of rules, that explains the training data as good as possible. [TODO: Add reference].

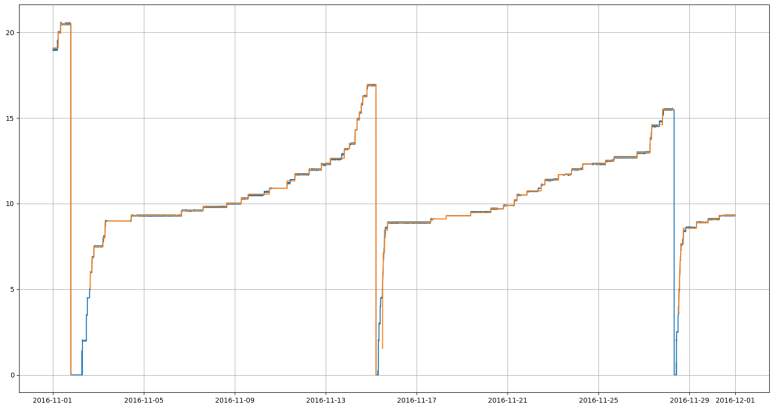

Figure 11: Discretization of Oven Power for November 2016. In blue the original data is plotted, the orange line represents the discrete approximation.

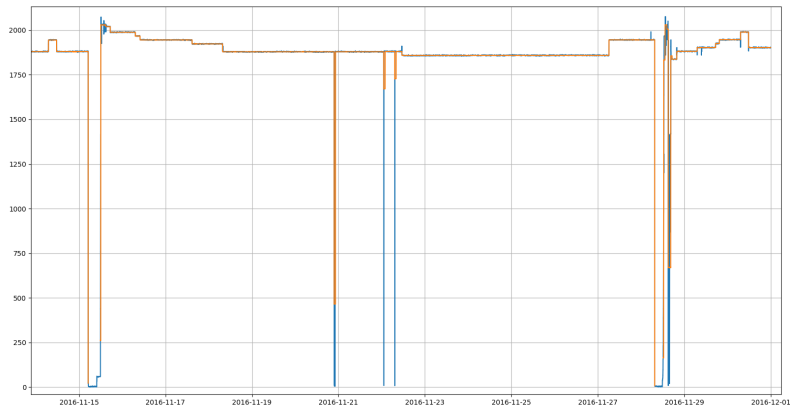

Figure 12: Discretization of RF Power for November 2016.

Decision trees are a supervised learning method, meaning that for each training input a output class is specified, and the algorithm tries to learn this relationship. Decision trees can be used for regression. If you want to regress a function $\mathbb{R}^n \to \mathbb{R}; \quad (x_1, \ldots, x_n) \mapsto y$ you pass $(x_1, \ldots, x_n)$ as input and y as the desired class. In the case of an one dimensional function for example, a decision tree classifier could lern that for $x >= 5$ and $x <= 10$ it should output $y = 5$. So, by the nature of a decision tree, the regression result is a step function, that looks like the result as much as possible.

One very common problem with decision trees is over fitting. If they are allowed to grow too much, they are not regressing any more, but copying. For example suppose that all your input data points are the natural numbers. By building a tree with the rules $(x >= 0.5, x < 1.5)$, $(x >= 1.5, x < 2.5)$, $(x >= 2.5, x < 3.5)$, ... the resulting regressor could perfectly replicate the input function, but it would learn all small oscillations, what is not what we are typically interested in. However, one thing that can be controlled is how many leaf nodes the tree can have, i.e. in the 1D case into how many intervals the real axis can be split at most.

In our case it would hence be useful to know how many discrete levels, or number of constant segments, the function we want to model consist of. Then, we could regress it using a tree with this maximum of leaf node, because by minimizing the error (i.e. difference from the original function) it would find the best stepwise approximation to the input data without over-fitting.

Finding the number of constant segments

Change Point Detection

Change points are abrupt changes in time series data that can for example be an indicator of some change in the underlying generating process. Finding change point can be useful in a lot of cases when it is interesting to calculate exact times when the nature of the data changed. On Linac3 example of a change point could be a sudden drop in the beam intensity. This could be an indicator of a change of some source parameter, and detecting such a change point early could be used to alert the operators.

There is a variety of change point detection Literature available, one overview can be found in [2]. Generally, one distinguishes between offline (or batch) and online analysis. In the offline setting, all data is available a priory, while in the online setting the algorithm decides if a change point occurred every time a new data point arrives.

Typically there are two approaches to change point detection: Either, one specifies some cost function that is to be minimized. This could be for example the L2-Distance when the number of change points is known an changes in mean should be detected (i.e. signal is a step function possibly with some noise and one wants to see when there was a jump). If the number of change points is not known, one needs to introduce a penalty that somehow depends on the number of change points, to avoid overfitting. For different cost functions in the offline setting and what kind of changes they capture see [10]. The second approach is to view a change point as a point, where some parameters of a probability distribution (or the distribution itself) change. Then one would do a hypothesis test with the H_0 hypothesis that the distributions before and after a given point x are equal. If the H_0

hypothesis is rejected, then there is a change point at x. A method which builds on this probabilistic setting is detailed in below.

Bayesian Online Change Point Detection

Bayesian Statistics

Bayesian statistics is a branch of statistics which makes use of Bayes' theorem to update its beliefs.

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

In conventional statistics, one has n observations of a random variable $x = (x_1, \dots x_n)$ coming from some probability distribution with unknown parameter $\boldsymbol{\theta}$ (which can be multidimensional, e.g. $\boldsymbol{\theta} = (\mu, \sigma^2)$ for the normal distribution). This parameter is treated as an unknown, but fixed variable. In Bayesian statistics however, the parameter is treated as well, so we can give meaning to $P(\boldsymbol{\theta} \mid x)$. It is the probability of $\boldsymbol{\theta}$ being the parameter of the distribution of the x_i, given all our observations. Hence, we have by Bayes theorem

$$P(\boldsymbol{\theta} \mid x) = \frac{P(x \mid \boldsymbol{\theta})\, P(\boldsymbol{\theta})}{P(x)} \propto P(x \mid \boldsymbol{\theta})\, P(\boldsymbol{\theta}) \tag{1}$$

$P(x \mid \boldsymbol{\theta})$ is called the likelihood, and it encodes our model of the reality, in how we think that x depend on $\boldsymbol{\theta}$. For the (stable) beam current we could for example assume that it is normally distributed with some parameters μ, σ were we estimate these values by looking at data from the last hour. $P(\boldsymbol{\theta})$ is called the prior distribution. It contains the knowledge we have about our parameter $\boldsymbol{\theta}$, by telling us, how likely a parameter is. For example an average beam current in the order of 1A is very unlikely, whereas an average current around 100mA is much more likely). Finally, $P(\boldsymbol{\theta} \mid x)$ is called the posterior distribution. It can be seen as an update of our observations on $\boldsymbol{\theta}$ with new knowledge coming from x. For a more thorough introduction see [9, 3].

Algorithm

Based on this idea, Adams and MacKay proposed an online change point detection algorithm [1]. They introduce the concept of *run lengths*. At a given point in time t the run length r_t gives the number of time steps that passed since the last change point happened. During one run, we assume that all observations come from the same distribution.

If the run length is known, at $t+1$ only two things are possible: $r_{t+1} = r_t + 1$ (no change point happened) or $r_{t+1} = 0$ (there was a change point). The algorithm computes at each step t the run length distribution $P(r_t \mid x_1, \dots, x_t)$, i.e. the probability of having a run length r_t having observed everything that happened up to time t, see figure 13.

By definition of the conditional probability we have $(x_1, \dots, x_t \equiv \boldsymbol{x}_{1:t})$

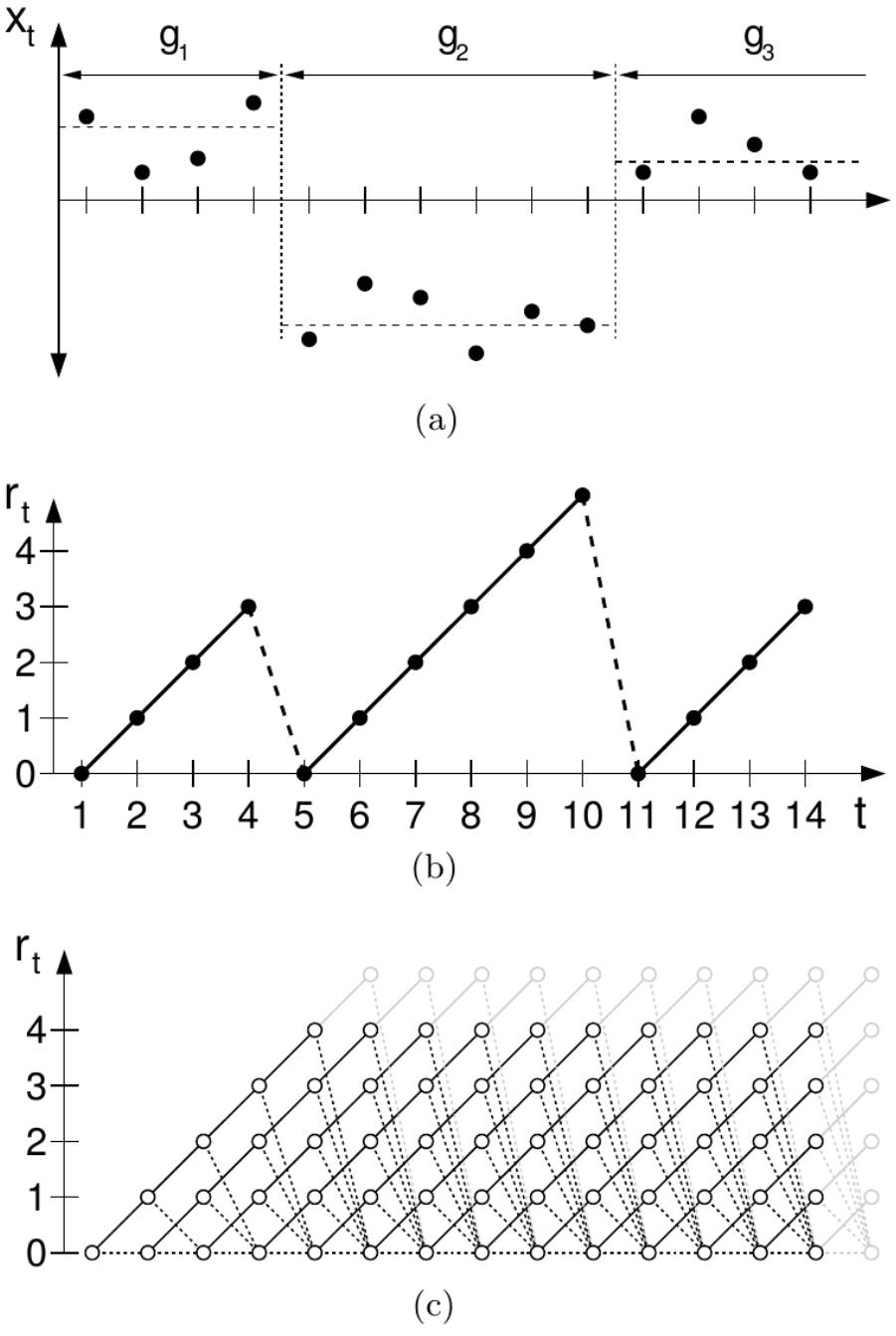

Figure 13: Explanation of the run length, figure taken from [1]. In (a) you can see a time series with two change points at $t = 5$ and $t = 11$. In (b) the true run length can be seen. Notice that when a change point occurs, the run length drops to zero. (c) shows all run lengths that are possible at a given time t. The algorithm introduced in [1] computes the probability of each node in the graph.

$$P(r_t \mid \boldsymbol{x}_{1:t}) = \frac{P(r_t, \boldsymbol{x}_{1:t})}{P(\boldsymbol{x}_{1:t})}$$

and the nominator can be computed recursively (see equation (3) in [1]) as

$$P(r_t, x_{1:t}) = \sum_{r_{t-1}} P(r_t \mid r_{t-1}) P(x_t \mid r_{t-1}, x_t^{(r)}) P(r_{t-1}, x_{1:t-1}) \tag{2}$$

while the denominator can be seen as a scaling factor. $x_t^{(r)}$ denotes the set of observations associated with the run r_t. To apply the algorithm it helps to understand some technicalities, which I try to explain below. However I recommend reading a little bit (for example on wikipedia) about Hazard functions and more importantly conjugate priors.

Hazard function

The first term in the sum, $P(r_t \mid r_{t-1})$, is called the change point prior and it encodes the information or intuition we have about the distribution of the change points. At any given time t, we can have only $r_t = r_{t-1} + 1$ or $r_t = 0$. So something called a hazard function H is introduced, that encodes the probability that a change point happened (The name hazard is because one can model like this the breaking probability of a mechanic piece which will probably depend on its lifetime). Then one can write

$$P(r_t \mid r_{t-1}) = \begin{cases} H(r_{t-1}+1) & \text{if } r_t = 0, \\ 1 - H(r_{t-1}+1) & \text{if } r_t = r_{t-1}+1 \text{ (no changpoint)}, \\ 0, & \text{else} \end{cases}$$

For the ion source it probably makes sense to model the change points distribution as an exponential distribution, meaning that at each time step there is a change of probability p for a new regime to start. However, one could improve this by making the hazard function dependent for example on a rising on the HT current, as a rising current seams to make drops in intensity more likely. But for the easy case without any assumption, only a sensible value for p needs to be determined. As for the exponential distribution we have $\mu = \frac{1}{p}$, one can choose p as one over the average time between change points, e.g. 4 hours. The algorithm doesn't seem to be to be very sensible to the exact value of p, as long as it is chosen reasonably. In this exponential case, the hazard function is just $H(y) \equiv p$, so it is constant.

Conjugate Priors

The first term in the sum of equation (2), $P(x_t \mid r_{t-1}, x_t^{(r)})$, gives the probability of observing an new value x_t, given that we have run length r_{t-1} and that for this run length we observed the values $x_t^{(r)}$. To calculate this, we need to decide how our data is distributed during one run.

In case of the beam current, it makes sense to assume it is normally distributed with unknown mean μ and variance σ^2. Using the first few points of the run, we could estimate

these two parameters, call them $\bar{\mu}$ and $\bar{\sigma}^2$, and then say that for the whole run we have $P(x_t \mid r_{t-1}, x_t^{(r)}) \sim Normal(\bar{\mu}, \bar{\sigma}^2)$. This would be one very likely Normal distribution the data could come from, but as we only estimated the parameters, there is some uncertainty involved. Hence it would make sense to use a weighted average distribution of all possible parameters, and for this we can use equation (1).

$$\begin{aligned} P(x_t \mid r_{t-1}, x_t^{(r)}) &= \int_\theta P(x_t \mid \theta) P(\theta \mid r_{t-1}, x_t^{(r)}) d\theta \\ &= \int_\theta P(x_t \mid \theta) \frac{P(r_{t-1}, x_t^{(r)} \mid \theta) P(\theta)}{P(r_{t-1}, x_t^{(r)})} d\theta \end{aligned} \tag{3}$$

Now, the important this to understand here, is that this integral can be very difficult to calculate and can be typically only been simulated numerically. But, for some combinations of likelihoods and priors, the result has an algebraic representation making random sampling unnecessary. If this is the case, one calls the prior a *conjugate prior*. There can be found lists (eg. [4]) of such pairs and how the parameters need to be updated. It is important to still have in mind that the likelihood and prior should be chosen to reflect reality.

Here is a short example where the likelihood is a normal with unknown mean μ and variance σ^2, as this is used in many implementations of the Algorithm and makes sense for the ion source data. Assume that we observed $\nu = 100$ data points and calculated a mean of $\hat{\mu} = 10$ and a standard deviation of $\hat{\sigma}^2 = 0.1$. This means that the sum of squares $\beta := \sum_{i=1}^{100}(x_i - \hat{\mu})^2 = 100 * \hat{\sigma}^2 = 10$. We now need prior distributions for μ and σ^2 of our normal model. We can make the following assumption: $\hat{\sigma}^2 \sim InverseGamma(\frac{\nu}{2}, \beta)$ If one looks at a graph of the probability density function of the Inverse Gamma with the parameters 50 and 10, one can see the graph has a bump around 0.1, so we can actually interpret it as our uncertainty in the calculated variance!

From this, it follows that $(x, \hat{\sigma}^2) \sim NormalInverseGamma(\hat{\mu}, \nu, \frac{\nu}{2}, \beta)$ by the definition of the Normal-Inverse-Gamma distribution.

Suffix arrays

References

References

[1] Ryan Prescott Adams and David J.V. MacKay. "Bayesian Online Changepoint Detection". In: *arXiv* (2007). URL: https://arxiv.org/pdf/0710.3742.pdf.

[2] Samaneh Aminikhanghahi and Diane J. Cook. "A survey of methods for time series change point detection". In: *Knowledge and Information Systems* 51 (2017). URL: https://link.springer.com/content/pdf/10.1007/s10115-016-0987-z.pdf.

[3] Jonny Brooks-Bartlett. *Probability concepts explained: Bayesian inference for parameter estimation.* Online Article. accessed 02.07.2020. Jan. 2018. URL: https://towardsdatascience.com/probability-concepts-explained-bayesian-inference-for-parameter-estimation-90e8930e5348.

[4] Daniel Fink. *A Compendium of Conjugate Priors*. 1997. URL: https://pdfs.semanticscholar.org/0894/42f59e3f4afb920479a7115c3dc57fb14757.pdf.

[5] A. Hinneburg and D. A. Keim. "Optimal Grid-Clustering: Towards Breaking the Curse of Dimensionality in High-Dimensional Clustering". In: *Proceedings of the 25th International Conference on Very Large Databases* (1999), pp. 506–517. URL: https://kops.uni-konstanz.de/bitstream/handle/123456789/5790/vldb99.pdf.

[6] Eamonn Keogh. *The UCR Matrix Profile Page*. URL: https://www.cs.ucr.edu/~eamonn/MatrixProfile.html.

[7] Sean M. Law. "STUMPY: A Powerful and Scalable Python Library for Time Series Data Mining". In: *The Journal of Open Source Software* 4.39 (2019), p. 1504. URL: https://github.com/TDAmeritrade/stumpy/.

[8] Dieter De Paepe, Olivier Janssens, and Sofie Van Hoecke. "Eliminating Noise in the Matrix Profile". In: *ICPRAM*. 2019. URL: https://biblio.ugent.be/publication/8605188/file/8605190.pdf.

[9] D. Spiegelhalter and K. Rice. "Bayesian statistics". In: *Scholarpedia* 4.8 (2009). revision #185711, p. 5230. DOI: 10.4249/scholarpedia.5230.

[10] Charles Truong, Laurent Oudre, and Nicolas Vayatis. "Selective review of offline change point detection methods". In: *Signal Processing, 167:107299* (2020). URL: https://arxiv.org/pdf/1801.00718.

[11] Eric W Weisstein. *Step Function*. From MathWorld–A Wolfram Web Resource. URL: https://mathworld.wolfram.com/StepFunction.html.

[12] Chin-Chia Michael Yeh, Nickolas Kavantzas, and Eamonn Keogh. "Matrix Profile IV: Using Weakly Labeled Time Series to Predict Outcomes". In: *VLDB 2017* (2017). URL: https://www.cs.ucr.edu/~eamonn/WeaklyLabeledTimeSeries.pdf.

[13] Chin-Chia Michael Yeh, Nickolas Kavantzas, and Eamonn Keogh. "Matrix Profile VI: Meaningful Multidimensional Motif Discovery". In: *ICDM 2017* (2017). URL: http://www.cs.ucr.edu/%7Eeamonn/Motif_Discovery_ICDM.pdf.

[14] Chin-Chia Michael Yeh et al. "Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View that Includes Motifs, Discords and Shapelets". In: *IEEE ICDM 2016* (2016). URL: https://www.cs.ucr.edu/~eamonn/PID4481997_extend_Matrix%20Profile_I.pdf.