Assessment of Unsupervised Standard Pattern Recognition Methods for Industrial Energy Time Series

Nicole Ludwig, Simon Waczowicz, Ralf Mikut, Veit Hagenmeyer
Institute for Automation and Applied Informatics
Karlsruhe Institute of Technology
firstname.lastname@kit.edu

ABSTRACT

Finding and extracting standard patterns in energy time series is very important to many real-world applications. Hence, there exists a multitude of pattern recognition algorithms with a majority of them being supervised ones. The advantage of supervision is that it can easily be checked if the algorithm is performing well or not. However, if no labels are available, an unsupervised pattern search is necessary. This search is faced with the challenge of how to measure success. Thus the question arises, when is a found pattern – for example a motif or a mean cluster curve – really describing the standard behaviour of a process and not just some kind of irrelevant behaviour? The present paper introduces a new method to assess two methods – namely clustering and motif discovery – in their quest to find standard profiles in energy time series data from industrial processes. Although both methods share the same aim, the results are incongruent. This has profound implications for real-world applications.

CCS CONCEPTS

• **Computing methodologies** → **Cluster analysis**; **Motif discovery**;

KEYWORDS

pattern recognition, time series, motif discovery, clustering

ACM Reference Format:
Nicole Ludwig, Simon Waczowicz, Ralf Mikut, Veit Hagenmeyer. 2018. Assessment of Unsupervised Standard Pattern Recognition Methods for Industrial Energy Time Series. In *Proceedings of ACM eEnergy Conference (eEnergy'18).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

Many machine learning algorithms want to find interesting patterns in time series data, as the information hidden in these patterns is crucial to most real world applications. In the realm of energy time series, knowing patterns especially in the consumption of energy is crucial to planning power supply and to maintain the stability in the grid. Answering questions like when does a process start is easy with supervised machine learning algorithms. For example we can find the signatures of certain machines efficiently with Non-intrusive Load Monitoring (NILM) [2]. However, if we do not know what patterns we are looking for, the task becomes significantly harder.

The major reason for this added difficulty is that we hardly know when we have succeeded in finding correct patterns. For example, when we cluster time series, we need to choose a stopping criteria or an objective function according to which the clusters are formed. The resulting cluster then depend largely on what was chosen. Furthermore, using e. g. motif discovery leaves us with the choice of a proper distance function and several other input variables. Again, the results depend largely on those chosen values. Without supervision it is difficult to properly evaluate the results obtained with either method.

In the present paper, we examine two well-established algorithms, namely fuzzy c-means clustering and motif discovery, which intend to find patterns in time series data, in a new way. We let both methods evaluate energy time series from an industrial consumer, a small electronics factory. The quest for both is to establish *standard profiles* for the process data at hand and categorise each section of the data as belonging to such a standard profile or being noise. We are interested whether the two methods find the same results when there is no knowledge on what processes are to be expected in the time series.

2 METHODOLOGY

Having established the need for process identification methods and the reason we chose to compare motif discovery and clustering, two well established methods. More specifically we use the fuzzy c-means clustering as described in [4] and the motif mining algorithms as described in [1, 3].

We evaluate our approach with real world data from a factory producing electronics. The data gathered consists of the active power, which we will refer to as power from here onwards, in kW of four machines over a time period of roughly two years. The maximum power of the machines ranges from 0.4 kW to 15.6 kW, while the standard deviation ranges from 0.06 to 0.58kW.

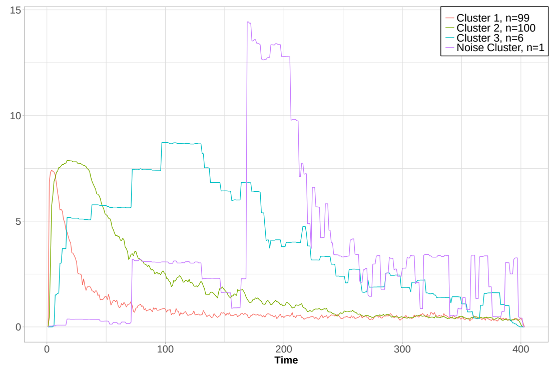
3 EVALUATION

The first step in our new evaluation method is concerned with the shape of the mean cluster and mean motif curves. Figure 1 shows the mean curves for the second machine.

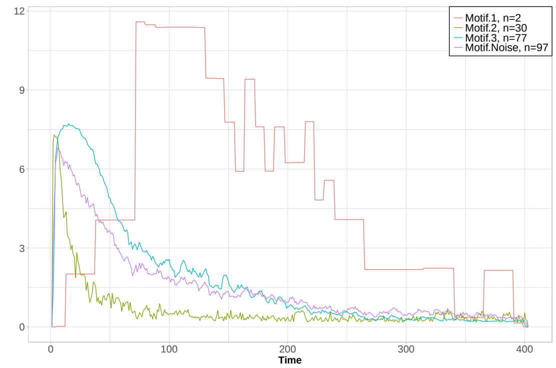
Evaluating the mean shapes,we can see two prominent differences. First, the magnitude of the value of the mean curves can vary substantially. Although both algorithms are working on the same data, the mean motifs are described by loads up to approximately

eEnergy'18, June 12 – 15, 2018, Karlsruhe, Germany

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4



(a) Cluster mean curves for machine 2.



(b) Motif mean curves for machine 2.

Figure 1: The mean curves for machine 2 of the cluster (a) in comparison to the mean curves of the motifs (b), with the number of subsequences associated with the respective mean. Note that the noise has not been classified as belonging to another cluster or motif, respectively.

0.6 kW, while the mean clusters range until almost 2 kW. The second distinction between the shapes is more relevant. Even though three out of four times the amount of groups found is the same, the number of sequences assigned to any cluster, motif or noise, vary considerably. Most prominently, the number of sequences classified as noise is higher for the motifs than for the clustering algorithm. While this is also due to parametrisation, we found that adjusting the motif discovery algorithm to find only few noise sequences resulted in the algorithm classifying all instances as being the same. However, this classification as one motif was only possible when allowing for vast distances between the sequences.

This discrepancy between how much noise is in the sequences is quite essential. Since a classification into another cluster means the algorithm finds a pattern connecting those sequences. However, the motif discovery algorithm does not find a strong enough relationship among those sequences. Overall, however, the shape of the mean curves are similar for both methods, indicating that on a coarse level the choice of methodology seems unimportant.

The second step in our evaluation leads us to the subsequence level. We want to know, whether the above-analysed shapes are determined with the same set of sequences or not. For this analysis, we examine to which motif the cluster subsequences were allocated. To do this, we take the cluster results and investigate which subsequences were categorised into which cluster. Given this cluster information, we now inspect to which motif the subsequences in each cluster were assigned by the motif discovery algorithm.

For example, for machine 1, the first cluster consists of 14 subsequences. According to the motif discovery, three of those sequences belong to motif 1, four belong to motif 2, and the remaining 7 are noise. Both algorithms do not agree on an assignment of sequences to the same groups. This result is fascinating, especially for applications where a precise allocation into groups is essential.

4 CONCLUSION & OUTLOOK

In the present paper, we have assessed exemplarily the results of unsupervised pattern recognition for industrial energy time series in a new way. Given the results from the two algorithms, we compare the patterns found and analyse which of the sequences are classified into the same group by both methods. We find that the mean shapes of the groups from both methods are comparable. However, we also see that the algorithms do not necessarily put the same sequences into those groups.

This incongruency between the two methods could have a tremendous impact and thus needs further investigation. If the results are profoundly different – as is the case in our example – then the choice of the method has far-reaching implications: Most applications use pattern recognition as the first step for a broader analysis, e. g., in order to build behavioural groups to target those groups individually. If the assignment into those groups is dependent on the algorithm, then an uneducated choice between the available methods distorts the results.

Hence, further research should analyse whether there are specific patterns, which one algorithm can undoubtedly find while the other cannot. Moreover, it is desirable to establish which algorithm is preferable for which specific application.

ACKNOWLEDGMENTS

This work was supported by the German Research Foundation (DFG) as part of the Research Training Group GRK 2153: Energy Status Data - Informatics Methods for its Collection, Analysis and Exploitation.

REFERENCES

[1] Bill Chiu, Eamonn Keogh, and Stefano Lonardi. 2003. Probabilistic discovery of time series motifs. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, Ted Senator (Ed.). ACM, New York, NY, 493–498.

[2] George W. Hart. 1992. Nonintrusive appliance load monitoring. *Proc. IEEE* 80, 12 (1992), 1870–1891. https://doi.org/10.1109/5.192069

[3] Nicole Ludwig, Simon Waczowicz, Ralf Mikut, and Veit Hagenmeyer. 2017. Mining Flexibility Patterns in Energy Time Series from Industrial Processes. In *Proceedings. 27. Workshop Computational Intelligence, Dortmund, 23. - 24. November 2017*, Frank Hoffmann, E. Hüllermeier, and Ralf Mikut (Eds.). KIT Scientific Publishing, 13–32.

[4] Simon Waczowicz, Markus Reischl, Veit Hagenmeyer, Ralf Mikut, Stefan Klaiber, Peter Bretschneider, Irina Konotop, and Dirk Westermann. 2015. Demand response clustering - How do dynamic prices affect household electricity consumption?. In *2015 IEEE Eindhoven PowerTech*. IEEE, 1–6. https://doi.org/10.1109/PTC.2015.7232493