Demand Response clustering — Automatically finding optimal cluster hyper-parameter values

Simon Waczowicz, Nicole Ludwig, Jorge Á. G. Ordiano, Ralf Mikut, Veit Hagenmeyer
simon.waczowicz@kit.edu
Institute for Automation and Applied Informatics
Karlsruhe Institute of Technology

ABSTRACT

Time series clustering methods, such as *Fuzzy C-Means* (FCM) noise clustering, can be efficiently used to obtain typical price-influenced load profiles (TPILPs) through the data-driven analysis and modelling of the consumption behaviour of household electricity customers in response to price signals (*Demand Response*, DR). However, the analysis of load time series with cluster methods presupposes that the user has a lot of experience in selecting good cluster hyper-parameter values (e.g. number of clusters or fuzzifier). The present contribution proposes a practical method to the automatic selection of optimal hyper-parameter values for DR clustering.

KEYWORDS

data mining, clustering, hyper-parameter, demand response

1 INTRODUCTION

The use of the demand side flexibility via control signals or price signals is a promising approach to balance demand and supply in smart grids [5]. The customers' consumption behavior can be analyzed and modelled with data mining methods, such as time series clustering [2]. The FCM clustering of smart meter data with the goal of extracting TPILPs is shown and tested in [8]. Furthermore, the authors of [8] suggest to apply noise clustering on FCM [6], because noise in smart meter data could significantly affect the found clusters. In addition to the goal of finding TPILPs, the DR clustering should reveal seasonal differences or differences in the consumption behavior on weekdays and weekends in response to a price signal. The decision, whether the found TPILPs are representing the typical consumption behavior of electricity customers in a meaningful way or not, presupposes that the users have a lot of experience both in the area of clustering and of the power sector. As an extension of [8], the new approach has the goal to automatically find and select optimal hyper-parameter values for DR clustering.

2 METHODOLOGY

Within the FCM noise clustering algorithm, the first step is to define the number of clusters C, the fuzzifier q and the noise distance δ. These cluster hyper-parameters are decisive for the cluster result. Hence, it is necessary to automatically find and select the best cluster hyper-parameter values for a given smart meter dataset[1].

The clustering is carried out multiple times, each time with a different combination of the values of the cluster hyper-parameters C, q and δ. The result of each clustering is evaluated using various cluster quality measures (Xie-Beni index S [9], global silhouette index SI [7]). The indexes indicate the similarity of data tuples within one cluster (cohesion) and the dissimilarity of a cluster from other clusters (separation). Additionally, the share of data tuples assigned to the noise cluster in relation to the total number of data tuples is used as a valuation criteria. This percentage share is described by the parameter τ_{oc}, where a low τ_{oc} corresponds to few data tuples and a high τ_{oc} corresponds to many data tuples in the noise cluster. The dependency between the found clusters and output variables, such as month, tariff or type of day, is quantified by the relative mutual information $Q_{(\mathrm{C,y})}$ [8].

We apply a simple but effective ranking method (PROMETHEE: Preference Ranking Organization Method for Enrichment Evaluations [1]) to compare each combination of hyper-parameter values. PROMETHEE performs a complete ranking of alternatives according to several valuation criteria in an automated manner.

3 DATA & RESULTS

The analyzed smart meter dataset comes from the *Olympic Peninsula Project*[2], in which the consumption behavior of electricity customers in response to variable electricity tariffs and suitable technology or bi-directional communication was examined [3].

We perform a comparison of seven strategies to find optimal cluster hyper-parameter values. The mean cluster curves are presented in Figure 1. The first strategy is an excerpt from [8], where the values for the hyper-parameters q and δ are defined by *manual parameter tuning*. Afterwards, the optimal value for the hyper-parameter C is derived by the first local minimum of S with $S(C)$

[1]The MATLAB toolbox SciXMiner [4] is used for the automatic selection of the hyper-parameter values.
[2]Download of the dataset with previous registration under https://svn.pnl.gov/olypen

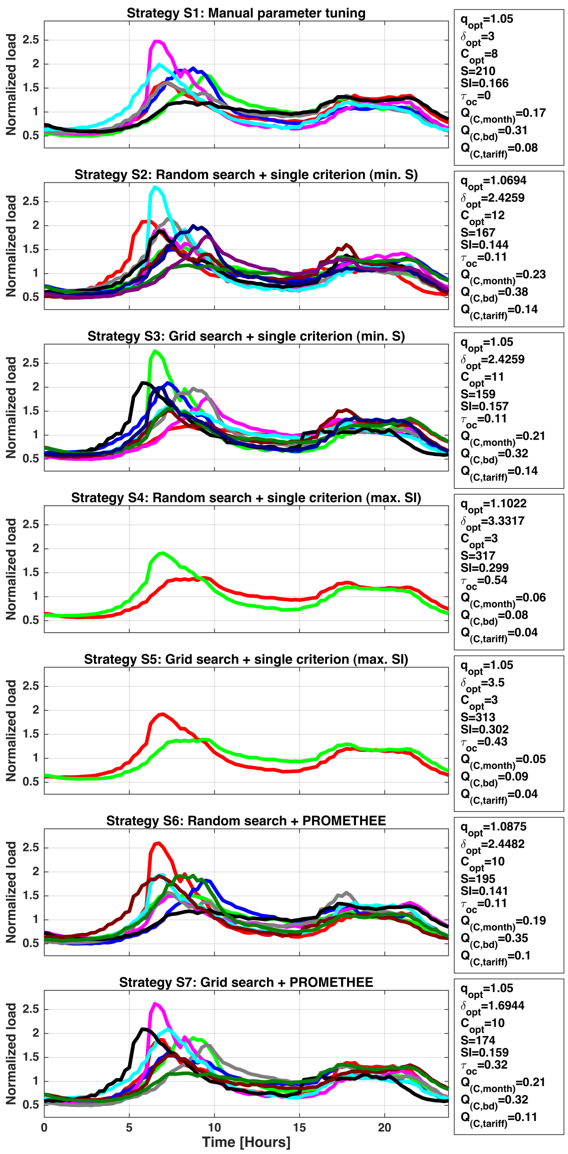

Figure 1: Clusters found for the aggregated, normalized daily time series of *Olympic Peninsula Project* dataset.

and $C = 3 \ldots 15$. The strategies S2, S4 and S6 use a *random search* approach with N_{CP} repetitions to set the values for the hyper-parameters C, q and δ. In case of S2 and S4, the optimal values for the hyper-parameters are derived by the minimum of the Xie-Beni index S with $S(q, \delta, C)$ and by the maximum of the global silhouette index SI with $SI(q, \delta, C)$, respectively. S6 is a combination of *random search* and PROMETHEE. S3, S5 and S7 correspond to S2, S4 and S6, but apply *grid search* instead of *random search*. S7 is a combination of *grid search* and PROMETHEE.

The results indicate that the strategy for choosing optimal hyper-parameter values plays a decisive role regarding the number of clusters. However, the optimal values for the hyper-parameters q and δ are within a small value range for all strategies. Furthermore, the method for setting the hyper-parameter values is not decisive. By randomly setting the values for the hyper-parameters C, q and δ and performing the clustering 1000 times, the whole value ranges of C, q and δ are well covered. The method for ranking the cluster runs determines the values of the hyper-parameters. If the Xie-Beni index S is used as the only ranking criterion (S2, S3) there are many cluster centers. The ranking of the cluster runs based on SI (S4, S5) leads to few cluster centers.

The relative high values for $Q_{(\mathrm{C,month})}$ and $Q_{(\mathrm{C,bd})}$ (index bd: business day, yes/no) for S6 and S7 indicate that the seasonal and weekday influence on the price-influenced consumption behavior can be explained by the found clusters. The comparison of S1 and S7 with respect to the $Q_{(\mathrm{C,y})}$ values shows that the S7 clusters are better suited to explain the seasonal and weekday influence price-influenced consumption behavior. In addition, with S7 (and S6) the time-intensive search for optimal hyper-parameter values is no longer necessary. Another advantage of using the PROMETHEE method is that the ranking also allows you to consider the second and third best solutions.

4 CONCLUSIONS

We introduce a new strategy to automatically find optimal hyper-parameter values for DR clustering. The strategy consists of two parts: (1) a comprehensive time series clustering with different hyper-parameter values (*grid search*) and (2) an effective ranking method (PROMETHEE) to compare and assess the cluster runs. We compare the new strategy (S7) with six other strategies (S1-S6) and are able to show that the clusters found by S7 are better suited to explain the seasonal and weekday influence on the price-influenced consumption behavior than the other strategies.

The great advantage of the new strategy is that it provides optimal hyper-parameter values for DR clustering, taking into account several valuation criteria. The use of the new strategy also enables non-specialists to evaluate the result of a time series clustering and to find optimal combinations of hyper-parameter values according to the selected valuation criteria and the weighting of these criteria.

REFERENCES

[1] J.-P. Brans and P. Vincke. 1985. A Preference Ranking Organisation Method (The PROMETHEE Method for Multiple Criteria Decision-Making). *Management science* 31, 6 (1985), 647–656.
[2] S. Haben, C. Singleton, and P. Grindrod. 2016. Analysis and Clustering of Residential Customers Energy Behavioral Demand Using Smart Meter Data. *IEEE Trans. Smart Grid* 7, 1 (2016), 136–144.
[3] D. J. Hammerstrom, R. Ambrosio, J. Brous, T. A. Carlon, D. P. Chassin, et al. 2007. *Pacific Northwest GridWise Testbed Demonstration Projects, Part I. Olympic Peninsula Project.* Technical Report. Pacific Northwest National Laboratory (PNNL).
[4] R. Mikut, A. Bartschat, W. Doneit, J. Á. González Ordiano, B. Schott, J. Stegmaier, S. Waczowicz, and M. Reischl. 2017. *The MATLAB Toolbox SciXMiner: User's Manual and Programmer's Guide.* Technical Report. arXiv:1704.03298.
[5] P. Palensky and D. Dietrich. 2011. Demand Side Management: Demand Response, Intelligent Energy Systems, and Smart Loads. *IEEE Trans. Industr. Inform.* 7, 3 (2011), 381–388.
[6] F. Rehm, F. Klawonn, and R. Kruse. 2007. A Novel Approach to Noise Clustering for Outlier Detection. *Soft Computing* 11, 5 (2007), 489–494.
[7] P.J. Rousseeuw. 1987. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *J. Comput. Appl. Math.* 20 (1987), 53–65.
[8] S. Waczowicz, M. Reischl, V. Hagenmeyer, R. Mikut, S. Klaiber, P. Bretschneider, I. Konotop, and D. Westermann. 2015. Demand Response Clustering - How do Dynamic Prices Affect Household Electricity Consumption?. In *Proc., IEEE PowerTech.* IEEE, Piscataway, NJ, USA, 1–6.
[9] X.L. Xie and G. Beni. 1991. A Validity Measure for Fuzzy Clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* 13, 8 (1991), 841–847.