District Heating Substation Behaviour Modelling for Annotating the Performance

Shahrooz Abghari[1(✉)], Veselka Boeva[1], Jens Brage[2], and Christian Johansson[2]

[1] Blekinge Institute of Technology, 371 79 Karlskrona, Sweden
shahrooz.abghari@bth.se
[2] NODA Intelligent Systems AB, 374 35 Karlshamn, Sweden

Abstract. In this ongoing study, we propose a higher order data mining approach for modelling district heating (DH) substations' behaviour and linking operational behaviour representative profiles with different performance indicators. We initially create substation's operational behaviour models by extracting weekly patterns and clustering them into groups of similar patterns. The built models are further analyzed and integrated into an overall substation model by applying consensus clustering. The different operational behaviour profiles represented by the exemplars of the consensus clustering model are then linked to performance indicators. The labelled behaviour profiles are deployed over the whole heating season to derive diverse insights about the substation's performance. The results show that the proposed method can be used for modelling, analyzing and understanding the deviating and suboptimal DH substation's behaviours.

Keywords: Clustering analysis · District heating · Higher order mining · Outlier detection

1 Introduction

A district heating (DH) system provides an entire town, or part of it, with heat. The heat is generated in a central boiler and delivered via a distribution pipe network. The provided heat transfers through DH substations from the distribution network into consumers' buildings. This includes providing both space heating for heating seasons and domestic hot water (DHW) for a whole year. The DH system consists of two sides: *primary* and *secondary*. The primary side includes a central boiler, a distribution network (pre-insulated pipes) and consumers' buildings. The secondary side consists of a heat exchanger, a main piping system of the building, and radiators, convectors, or floor heating for the rooms.

This work is part of the research project "*Scalable resource-efficient systems for big data analytics*" funded by the Knowledge Foundation (grant: 20140032) in Sweden.

P. Cellier and K. Driessens (Eds.): ECML PKDD 2019 Workshops, CCIS 1168, pp. 3–11, 2020.
https://doi.org/10.1007/978-3-030-43887-6_1

The DH substations are made up of different components and each can be a potential source of faults. Faults in substations and the secondary side can be divided into three categories (1) faults resulting in comfort problems such as lack of enough heat, (2) unsolved faults with known cause since their identification are time demanding and costly, and (3) faults that require advanced fault detection systems [1]. Faults in substations do not necessarily result in comfort problems for the consumers, instead in most cases cause sub-optimal behaviour for a long time before they are noticed. Therefore, early detection of faults and deviations can reduce the maintenance cost and help avoid abnormal event progression. Fault detection in DH substations can be performed by monitoring both primary and secondary sides or only primary side.

Gadd and Werner [1] showed that hourly meter readings can be used for detecting faults at DH substations. The authors identified three fault groups: (1) low average annual temperature difference, (2) poor substation control, and (3) unsuitable heat load patterns. The results of the study showed that addressing low average annual temperature differences are the most important issue that can improve efficiency of the DH systems. Nevertheless, unsuitable heat load patterns are probably the easiest and the most cost-effective problem to consider first. In a recent study [2], the authors applied clustering analysis and association rule mining to detect faults in DH substations. In another study, the authors [3] proposed a method based on gradient boosting regression to predict hourly mass flow of a well performing substation. Their built model was tested by manipulating well performed substation data to simulate two different scenarios. Calikus et al. [4] proposed an approach to automatically discover heat load patterns in DH systems. Heat load profiles reflected yearly heat usage in an individual building. Moreover, their discovery is crucial for ensuring effective DH operations and managements.

We propose a higher order mining (HOM)[1] approach for modelling a DH substation's operational behaviour and linking it with two performance indicators. At the modelling step, we use primary side features to build the substation behaviour model by extracting the substation's behaviour patterns on a weekly basis. Heat demand is strongly influenced by social factors, e.g., the need during weekdays versus weekends. However, the social patterns tend to repeat on a weekly basis. Therefore, by considering the time window of a week rather than a day, we can mitigate the social patterns and avoid discovering, e.g., the demand transition between weekdays and weekends. The extracted patterns are used to create weekly behaviour models by clustering them into groups of similar patterns. The built models are further analyzed and integrated into an overall substation model by applying consensus clustering. We consider the exemplars of the consensus clustering model as the substation representative operational behaviour profiles. Further, at the annotating step the exemplars are linked with the two performance indicators. These indicators are calculated by using features from both primary and secondary side data. The annotated behaviour profiles can be deployed over the whole heating season to derive diverse insights about

[1] HOM is a sub-field of knowledge discovery that applies to non-primary, derived data or patterns to provide human-consumable results [5].

the substation's performance. They can also be used to quantify the performance of incoming heating weeks.

2 Methods and Techniques

2.1 Sequential Pattern Mining

Sequential pattern mining is the process of finding frequently occurring patterns in a sequence dataset. The records of the sequence dataset contain sequences of events whose orders are important. We use the PrefixSpan algorithm [6] to extract frequent sequential patterns. PrefixSpan applies a prefix-projection method recursively to find sequential patterns. The prefix-based projection enables PrefixSpan to focus only on prefix sub-sequences and project on their corresponding postfix sub-sequences. This yields less projections which in turn reduces both the length and the number of sequences in the projected datasets.

2.2 Clustering Analysis

Affinity Propagation: We use the affinity propagation (AP) algorithm [7] for clustering the extracted patterns. AP is based on the concept of *message passing* between data points. Unlike clustering algorithms, such as k-means [8] which requires the number of clusters as an input, AP estimates the optimal number of clusters from the data. In addition, the chosen exemplars are real data points and representative of the clusters.

Consensus Clustering: Gionis et al. [9] proposed an approach for clustering based on the concept of aggregation, where a number of different clustering solutions are given on some datasets of elements. The objective is to produce a single clustering solution from those elements that agrees as much as possible with the given clustering solutions. Consensus clustering algorithms deal with similar problems to those treated by clustering aggregation techniques. Such algorithms aim to synthesize clustering information about the same phenomenon coming from different sources [10] or from different runs of the same algorithm [11]. In this study, we use the consensus clustering schema proposed in [10] in order to integrate the clustering solutions produced on the datasets collected on a weekly basis for the heating season. The exemplars of the produced clustering solutions are considered and divided into k clusters according to the degree of their similarity by applying the AP algorithm. Subsequently, clusters whose exemplars belong to the same partition are merged in order to obtain the final consensus clustering.

2.3 Distance Measure

The similarity between the extracted patterns are assessed with a dynamic programming version of Levenshtein distance (LD) metric [12]. The LD, also known as edit distance, is a string similarity metric that measures the minimum number of editing operations required to transform one string into the other.

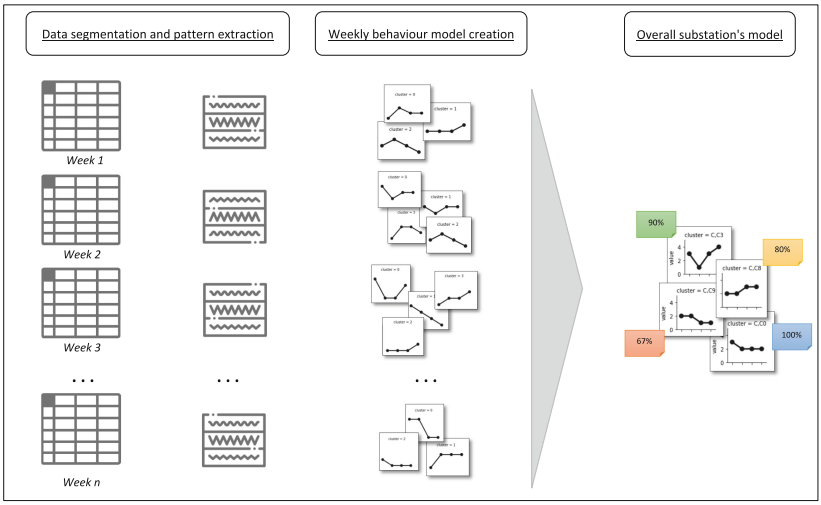


Fig. 1. Schematic illustration of the proposed approach

3 Proposed Method

Our approach has a preprocessing step and two main steps: (1) *Modelling substation's operational behaviour*; (2) *Linking the substation's representative behaviour profiles with performance indicators*. The modelling step consists of three distinctive sub-steps: (i) data segmentation and pattern extraction, (ii) weekly behaviour model creation, and (iii) overall substation's model. The approach is schematically illustrated in Fig. 1.

Data Preprocessing: In order to prepare data for the modelling step all duplicates are removed and missing values are imputed by averaging the neighbouring values. The first and the last missing values are replaced with the next and the previous available values, respectively.

In addition, extreme values that are often a result of faults in measurement tools are smoothed out by a Hampel filter [13], which is a median absolute deviation (MAD) based estimation. The filter computes the median, MAD, and the standard deviation (SD) over the data in a local window. We apply the filter with the default parameters; the size of the window is seven and the threshold for extreme value detection is three, i.e., 3-neighbours on either side of a sample. The threshold for extreme value detection is three. Therefore, in each window a sample with the distance three times the SD from its local median is considered as an extreme value and is replaced by the local median.

We monitor the operational behaviour of substations based on outdoor temperature and the primary side features of the DH system. Our motivation for this choice relates to the fact that the primary side data is always available while the secondary side data requires specific hardware that might not be available at the consumers' building. After discussions with domain experts, we chose

five features that have a strong negative correlation with outdoor temperature. The selected features are: (1) primary return temperature, $T_{r,1^{st}}$, (2) primary temperature difference, $\Delta T_{1^{st}}$, (3) primary energy, $Q_{1^{st}}$, (4) primary mass flow rate, $G_{1^{st}}$, and (5) the substation performance indicator based on the hourly consumed energy divided into the hourly mass flow rate, E_s^{E-F}. The fifth feature represents how many units of energy one substation can provide from the consumed volume flow rate.

Z-score normalization is applied on each feature and for every week's period. The normalization is performed to make it possible to assess and compare a substation's operational behaviours in different weeks.

In order to build the DH substation's operational behaviour model using the HOM paradigm, continuous features are converted into categorical features. All five features together build patterns (sequences of events) that represent the operational behaviour of the substation. In this study, we are interested in contextual outlier detection. The context here is referred to as modelling the DH substation's behaviour, during only the heating season. For this purpose we have applied k-means-based discretization method by setting the size of k to four, similar to the number of seasons in Sweden.

1. Modelling DH substation's operational behaviour:

(i) **Data segmentation and pattern extraction:** We extract the substation's behaviour patterns on a weekly basis. The PrefixSpan algorithm is used to find frequent sequential patterns with the length of five in each week. Those sequential patterns that satisfy the user-specified support are considered as frequent ones. The user-specified support threshold is set to be *one* to capture daily patterns, i.e., any patterns that appear at least once will be considered.

(ii) **Weekly behaviour model creation:** The extracted patterns from each week are clustered into groups based on their similarities. Since the aim is to build a DH substation behaviour model for the heating season, all exemplars of the clustering models related to the weeks with the average outdoor temperature above 10 °C are filtered out.

(iii) **Overall substation's model:** The weekly behaviour models built at the previous step are further integrated into an overall substation's behaviour model by applying a consensus clustering technique. The exemplars of the consensus clustering solution are considered as representative profiles for the substation's behaviour, i.e., they can be used to further analyze the substation's behaviour and performance for the whole heating season.

2. Linking behaviour profiles with performance indicators: At this step the derived substation's behaviour profiles are linked to performance indicators. In the current study, we annotate behaviour profiles with two performance indicators: *substation effectiveness* and *grädigkeit*. The two indicators are computed by considering features from both the primary and secondary sides.

Substation effectiveness is computed as $E_s^T = \frac{\Delta T_{1^{st}}}{T_{s,1^{st}} - T_{r,2^{nd}}}$ where, $\Delta T_{1^{st}}$ is the difference between primary supply and return temperatures, $T_{s,1^{st}}$ is the

primary supply temperature, and $T_{r,2^{nd}}$ is the return temperature at the secondary side. The efficiency of a well-performed substation should be close to 1 in a normal setting. However, due to the affect of DHW generation on the primary return temperature, the E_s^T can be above 1.

Grädigkeit indicator, also known as the least temperature difference[2], represents the difference between primary and secondary return temperatures and it is computed as $\Delta T_{r,(1^{st},2^{nd})} = T_{r,1^{st}} - T_{r,2^{nd}}$. The grädigkeit of a substation can be greater than or equal to zero, though it can go below zero due to usage of DHW. A lower value of grädigkeit implies better performance.

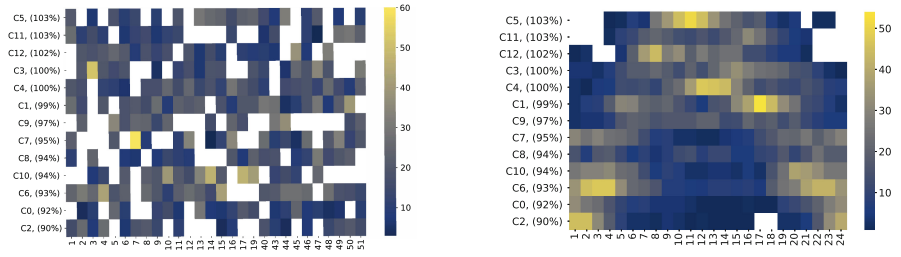
For each considered performance indicator, we partition the substation's representative behaviour profiles into three categories with respect to the associated performance indicator scores: *low*, *medium* and *high*. In that way, we have a group of behaviour profiles that represents the substation's sub-optimal performance and two groups of profiles that are linked with satisfactory and optimal substation's performance, respectively. The labelled behaviour profiles can be deployed over the whole heating season in order to further analyze and understand the substation's operational behavior and performance. For example, the profiles from the three different categories can be used to interpret the substation's operational behaviours for particular time intervals. In addition, it is possible to backtrack from these higher order representative profiles to the weekly behaviour models and to the hourly patterns.

4 Results and Discussion

We studied substations' operational behaviour for ten buildings in 2017. We first modeled each substation's weekly operational behaviours. This was performed by grouping the extracted frequent patterns into clusters of similar patterns. We then stored the exemplars of the built clustering model if the average outdoor temperature of the week was less than or equal to 10 °C. This step is motivated by the fact that we want to model the substation's overall operational performance for the whole heating season. The collected exemplars were integrated into a consensus clustering. At last, the obtained consensus clustering model was linked (annotated) with the selected performance indicators. The extracted profiles with respect to each indicator were used to assess behaviour of the substation on a weekly basis.

For the rest of this section we focus on one specific building, B-21. We identified 13 profiles that model the operational behaviour of the substation for the heating season. The extracted profiles were linked with the two performance indicators, *substation effectiveness* and *grädigkeit*. In order to facilitate further analysis, the profiles were sorted from the highest to the lowest performance separately for each indicator. For example, in case of the substation effectiveness the profiles are within a range from 103% to 90%. Regarding the grädigkeit, the profiles are within a range from −2.15 °C to 5.37 °C.

[2] Frederiksen, S., Werner, S.: District heating and cooling, Studentlitteratur Lund (2013).



(a) Heatmap represents profiles' frequency in each week.

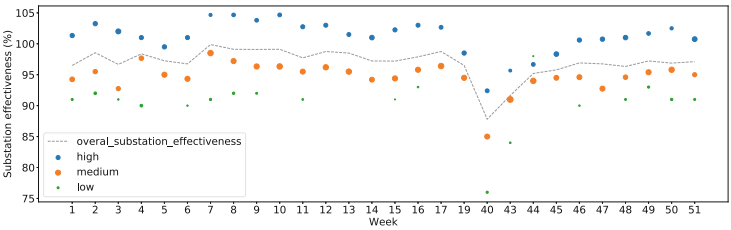
(b) Heatmap represents profiles' frequency in 24-hour period.

Fig. 2. The deployment of the annotated profiles according to *substation effectiveness* for building B-21 over 2017 heating season. (Color figure online)

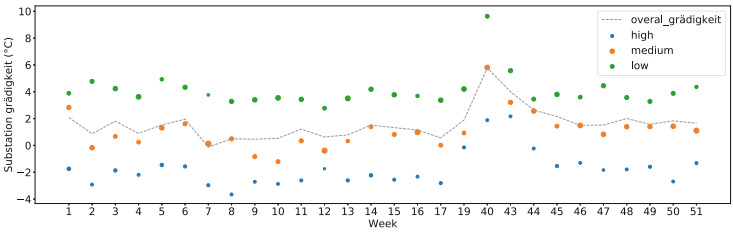
Figure 2a shows the substation's effectiveness according to the built profiles for each week. As one can notice the heatmap is sparse and only few weeks, e.g., weeks 3, 4, 7, 10, 14, 15, 17, and 18 represent a high number of frequency for some of the profiles. The heatmap is not easy to interpret and it does not provide interesting information about the substation's weekly behaviour. Figure 2b, on the other hand, provides more information by showing the effectiveness of the same substation at a 24-h period for the whole heating season. For example, one can recognize a yellowish bell shape. Evidently, the substation performed on average 92% at early morning (0:00–5:00) and late evening (20:00–23:00). However, for the rest of the day the performance of the substation is closer to and above 100%. The low performance of the substation might be due to social behaviour, which demonstrates low heat demand in the early morning and late evening.

As mentioned before, we categorize the extracted profiles with respect to their performance indicator labels (substation effectiveness or grädigkeit) into three categories: low, medium, and high. In the case of substation effectiveness *low* represents efficiency below 90%, *medium* indicates efficiency between 90% to 100%, and high stands for efficiency above 100%. Figure 3a shows the overall effectiveness of the substation over the weeks that space heating was required, based on these three categories. As one can see the orange circles, which represent the medium efficiency of the substation, closely follow the curve showing overall substation's effectiveness. This is also valid for the profiles from the other two categories. For example, all profiles linked to optimal performance (blue circles in Fig. 3a) are above the overall substation's effectiveness curve. In Fig. 3a, we can also notice that weeks 19 and 40 represent the end and beginning of the heating season, respectively. The low efficiency of the substation in week 40 might be related to the fact the system required sometime to adjust.

Regarding grädigkeit indicator, *low* represents temperature differences above 3 °C, *medium* denotes temperatures between 0 to 3 °C, and high shows temperature differences equal or below to 0 °C. Figure 3b shows the overall grädigkeit for the studied substation. Similar to Fig. 3a, the medium category is closely



(a) Substation's effectiveness.



(b) Substation's grädigkeit.

Fig. 3. The deployment of the annotated profiles according to performance indicators for building B-21 over 2017 heating season. (Color figure online)

following the curve that represents the overall substation's grädigkeit. Notice that for grädigkeit indicator the temperature differences close to and below zero show a high efficiency.

5 Conclusion and Future Work

We proposed a higher order mining approach for modelling a district heating substation's operational behaviour. The method summarized the substation's behaviour with a series of representative profiles that were linked with two performance indicators. The labelled profiles were deployed over the whole heating season to assess an overall substation's behavior and performance. We applied and studied our method on ten buildings. The initial results showed that the proposed method can be used to analyze and evaluate the operational behaviour of DH substations.

For future work we are interested in studying whether the derived representative behaviour profiles can be used to quantify the performance of incoming heating weeks. In addition we plan to evaluate our approach with other performance indicators.

References

1. Gadd, H., Werner, S.: Fault detection in district heating substations. Appl. Energy **157**, 51–59 (2015)
2. Xue, P., et al.: Fault detection and operation optimization in district heating substations based on data mining techniques. Appl. Energy **205**, 926–940 (2017)
3. Månsson, S., Kallioniemi, P.O.J., Sernhed, K., Thern, M.: A machine learning approach to fault detection in district heating substations. Energy Procedia **149**, 226–235 (2018)
4. Calikus, E., Nowaczyk, S., Sant'Anna, A., Gadd, H., Werner, S.: A data-driven approach for discovery of heat load patterns in district heating. arXiv preprint arXiv:1901.04863 (2019)
5. Roddick, J.F., Spiliopoulou, M., Lister, D., Ceglar, A.: Higher order mining. ACM SIGKDD Explor. Newsl. **10**(1), 5–17 (2008)
6. Pei, J., et al.: PrefixSpan: mining sequential patterns efficiently by prefix-projected pattern growth. In: Proceedings 17th International Conference on Data Engineering, pp. 215–224 (2001)
7. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. Science **315**(5814), 972–976 (2007)
8. MacQueen, J., et al.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Oakland, CA, USA, Vol. 1, pp. 281–297 (1967)
9. Gionis, A., Mannila, H., Tsaparas, P.: Clustering aggregation. ACM Trans. Knowl. Disc. Data **1**(1), 4-es (2007). https://doi.org/10.1145/1217299.1217303
10. Boeva, V., Tsiporkova, E., Kostadinova, E.: Analysis of multiple DNA microarray datasets. In: Kasabov, N. (ed.) Springer Handbook of Bio-/Neuroinformatics, pp. 223–234. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-642-30574-0_14
11. Goder, A., Filkov, V.: Consensus clustering algorithms: comparison and refinement. In: ALENEX, pp. 109–234 (2008)
12. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. Sov. Phys. Dokl. **10**, 707–710 (1966)
13. Hampel, F.R.: A general qualitative definition of robustness. Ann. Math. Stat. **42**, 1887–1896 (1971)