Using Sequence Mining to Understand Daily Activity Patterns for Load Forecasting Enhancement

Yong DING, Julio BORGES, Martin A. NEUMANN and Michael BEIGL
TECO, Institute of Telematics
Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany
Email: {firstname.surname}@kit.edu

Abstract—Load forecasting at appliance-level or house-level is a key to develop efficient Demand Side Management programs. Lots of recent research work have pointed out that load curves at household’s level depend highly on human behaviors and activities. However, the state-of-the-art load modeling approach takes only individual human activities with appliance-level time-of-use data into account. There is little study about influence of sequences of activities performed throughout a day on power consumption at household’s level. In this work, we conduct a broad study of activity sequences in daily life that influence power consumption of individual households. A context-rich data set including daily activity information and power consumption measurements from 24 households is collected across Japan. The contribution of this paper is 2-fold: 1) a set of insights into household-specific activity sequences influencing power consumption derived by a sequence mining algorithm, in order to identify significant associations between power consumption and household-specific activity sequences; 2) a load forecasting study using identified frequent activity sequences as an enhancement. Our analysis on sequence-based rules shows potential for inferring future activities and the power consumption of the future activity. Finally, we demonstrate how very short-term load forecasting, like 15 minutes ahead, can benefit from activity sequences for individual households.

I. INTRODUCTION

Dynamic human activity is an important contributor to the varying local power consumption, particularly in urban areas [1], [2], [3]. However, very few end-users fully understand the impact of their individual activities on power consumption and distribution [4], [5]. The key issues here include the lack of real-time information around consumption and also influence of human activity information on power consumption [6], [7]. Therefore, insights into human activities and their correlations with power consumption can help power suppliers to more accurately estimate the active demand, especially the short-term load, which is becoming particularly important due to the high fluctuations of the distributed renewable energy generation. At the same time, consumption anomaly prediction is vital for effective energy trading [8], as well as for grid stability, security and efficiency [9], [10].

Great efforts have been made towards extending human activities as primary influence variables in load forecasting models, in particular in time-of-use (TOU) based load models [11], [12], [13], [14], [15], [16], [17], [18]. However, the activity information that has been studied in the current research is either the high-level social activity [19] or the appliance-level usage activity [14]. In both cases, the activity information refers to an individual activity label. Beyond individual activity labels, Rollins and Banerjee have addressed in their recent work [20] associations that relate two usage activities of appliances, but without showing the benefit of these associations for load modeling or forecasting. The question arises whether we could efficiently utilize the collected common daily activity information to generate more complex activity associations in terms of sequences with diverse number of activities, in order to improve the load forecasting accuracy at household’s level.

Therefore, we focus in this work on researching the impact of activity sequences on load modeling and consumption pattern recognition. The contribution of this paper is 2-fold: 1) a set of insights into household-specific activity sequences based on collected daily activity labels by applying a sequence mining algorithm, in order to identify significant associations between power consumption and household-specific activity sequences; 2) a load forecasting study using identified activity sequences as an enhancement, which will be analyzed in comparison with activity alone as influencing factor.

The rest of the paper is organized as follows: a short introduction of the collected data set is provided in Sec. II. In Sec. III, we describe the applied sequence mining algorithm and discuss the household-specific activity sequences that not only represent their daily life patterns, but also indicate subsequently certain power consumption patterns by means of sequence-based rules. We then evaluate the house-level load forecasting with activity and sequence variables as influencing input variables for individual households in Sec. IV to show the contribution of identified activity sequences. Finally, Sec. V concludes the paper.

II. THE COLLECTED DATA SET

We prepared a time-of-use (TOU) data set of power consumption and activities to study the relationship between them. The activity information is presented in terms of daily actions of residents. Our data set refers to the daily activities and power consumptions (at 1-minute intervals) from 24 households for about 3 months.

Our experiment for collecting the data set was conducted from the beginning of December 2014 to the beginning of March 2015. The subjects were voluntary households from diverse places of Japan. They were asked to take a few minutes each day to input their daily activities. Due to smart meter downtimes and user engagements, in total, we were able to obtain 10,250 activity data (mainly from 23 households) and...
about 2,369,000 power consumption data (mainly from 21 households) by March 04, 2015.

The number of activity types were 52, which comprised indiscriminative activities such as “lunch” and “meal”, and also very personal activities. Therefore, we clustered them manually to 11 meta activities: $A = \{\text{“at home”, “bath”, “computer”, “cooking”, “eat”, “hobby”, “housework”, “media”, “out”, “relax”, “sleep”}\}$.

As mentioned in the introduction, our focus in this paper is on activity sequences rather than individual activities. The above 11 meta activities are not all applicable for all the households, which affects possible activity sequences for individual households. An activity sequence in our work is defined as a chronologically ordered list of activity labels, e.g. $S = \langle\text{cooking, eat, out}\rangle$. Since different households have varying sets of activity sequences, the relevance of each activity sequence $S$ can not be easily determined at a global level for all households. Therefore, the extraction of frequent activity sequences is conducted mainly for each household with various support values, which will be explained in detail in Sec. III.

Moreover, also the load curves of individual households of our experiment differ strongly from each other, see exemplary load curves in Figure 1.

![Fig. 1: Two representative households’ power consumption measured within the experiment duration](image)

III. SEQUENCE MINING TO UNDERSTAND DAILY ACTIVITY PATTERNS

Sequence Mining is a data mining technique that focuses on finding statistically relevant patterns in a sequence form for a given data set. Also in the research field of home energy management [20], sequence mining can be applied to capture the usage of appliances in sequence. Not only activities like the appliance usage alone can help understand how power consumption is influenced by certain activities, but also sequences of activities can deliver additional insights on how activities are related to each other. For example, the activity “out” after “eat”, can deliver the information, that this “eat” activity is rather related to a breakfast. Making a breakfast may require less energy consumption than cooking a dinner.

In this section, we conduct a sequence mining study of behavioral factors that influence power consumption in a household, identifying significant associations between power usage and sequences of activities performed by residents.

A. Household-specific Activity Sequences

A sequence is defined to be an ordered list of elements on a finite set $\zeta, S = \langle s_1, ..., s_n \rangle$, where each $s_i \in \zeta, i = 1..n$, $s_i$ appears in the position $i$ in $S$. The size of $S$ is referred by $n$ and denoted by $|S|$. A sequence $S_a = \langle a_1, ..., a_m \rangle$ is said to be contained in another sequence $S_b = \langle b_1, ..., b_n \rangle$, iff there exist integers $i_1 < i_2 < ... < i_m \leq n$, such that $a_1 = b_{i_1}, a_2 = b_{i_2}, ..., a_m = b_{i_m}$, denoted as $S_a \subseteq S_b$. $S_a$ is called the subsequence of $S_b$, and $S_b$ is called super-sequence of $S_a$.

The support of a sequence $S_j$ in a sequence dataset $SDB$ is defined as the portion of sequences $S \in SDB$, such that $S_j \subseteq S$. In our case, sequences of $SDB$ refer to sequences of activities performed throughout a day. Let $minsup$ be a threshold on the support value: a sequence $S_j$ is deemed frequent iff $support(S_j) \geq minsup$. The problem of mining sequential patterns is to discover all frequent sequences. For this task, different algorithms have been presented in the literature with distinct key features and techniques [21]. For our implementation in this work, we utilize the algorithm CM-SPADE [22] implemented in the SPMF library [23].

Within our collected data, it is possible to track sequences of different activities that the households perform throughout a day. To understand a household’s daily activity sequences, a sequence is given by all the activities performed by the household in a day, ordered by the start time of the activity. In this case, $\zeta = A$, i.e., the set of all possible activities as defined in II. For example, let $minsup = 0.5$, a sequence $S = \langle \text{cooking, eat, out} \rangle$ for a specific household is mined as frequent, it means that in at least 50% of the sequence occurrences when activity data for the household is available, the activity “out” has been performed in any time after the activity “eat”, and the activity “eat” has been performed in any time after the activity “cook”.

An overview of the number of the mined activity sequences is depicted in Figure 2 for the 23 households which annotated their activities frequently in our study. In comparison, the household with ID 72 exhibits rich activity sequences at each support level.

![Fig. 2: Overview of the number of frequent activity sequences from the 23 households with varying thresholds on the support value.](image)
Naturally, certain household presents activity patterns which are specific for this household and not shared among others. The activity sequence \( S_1 = \{\text{media, eat}\} \) for example is shared among 17 out of the 23 households, meaning that a high portion of the households perform media-related activity before eating. On the other hand, the activity sequence \( S_2 = \{\text{at home, cooking, eat, out, at home, cooking, eat, housework, out, at home, cooking, eat, sleep}\} \), describes the routine of the household with ID 72. In this household, it’s common to cook three times a day. This pattern is not shared among any other households.

Even among activity sequences shared among different households, the corresponding power consumption can vary significantly. For the previously discussed sequence \( S_1 = \{\text{media, eat}\} \), the household with ID 46 consumes in average 402.46 watts per minute \((\sigma = 52.60)\) whereas the household with ID 78 consumes in average 1933.67 watts per minute \((\sigma = 376.09)\) while both performing the same activity sequence. For this reason, we perform a sequence mining for household-specific activity sequences. Such sequence patterns are expected to describe the periodic daily routine of applicable households and improve their individual load forecasting.

**B. Inferring Power Consumption Influenced by Activity Sequences**

Different sequences can provide different context meanings for the same set of activities. For the sequence \( S_0 = \{\text{at home, cooking, eat, out}\} \), the activities are embedded in the context of a breakfast whereas in the sequence \( S_d = \{\text{out, at home, cooking, eat}\} \), the same activities (now in another sequence) are rather embedded in the context of a dinner. One would expect to consume more power for cooking a dinner than for making a breakfast. Our Analysis shows that this is exactly the case for households with these patterns. For instance, the household with ID 72 consumes in average only 309.94 watts per minute \((\sigma = 101.35)\) for the breakfast sequence, while consuming 454.07 watts per minute \((\sigma = 168.49)\) for the dinner sequence. The same set of activities can thus lead to different power consumption patterns when embedded in different contexts given by different sequences.

Motivated by this example, we are interested in inferring a future activity \( y \) and the power consumption while performing \( y \), given a previous observed sequence \( X \), i.e., \( \text{Power Load}(y) \mid X \). To solve this kind of problem, using sequence mining by mining frequent sequences alone is very limited. For example, consider the sequence pattern \( S = \{X, y\} \), which means that it is possible that \( y \) appears frequently after \( X \), but there may be also many cases where \( X \) is not followed by \( y \). For prediction, we need a measurement of the confidence that if \( X \) occurs, \( y \) will occur afterwards with a conditional probability \( P(y \mid X) \). In this case, we talk about a sequence rule, denoted by \( X \Rightarrow y \). The confidence is given by \( \text{confidence}(X \Rightarrow y) = P(y \mid X) = \frac{\text{support}(X \Rightarrow y)}{\text{support}(X)} \).

Using the activity information provided by the households can help us examine whether the sequence of certain activities in the household correlate with future activities and future power consumption. The time interval elapsed between the occurrence of \( X \) (i.e., the occurrence of the last element in \( X \)) and the occurrence of \( y \) is denoted as prediction horizon. Sequence mining can thus be leveraged for creating rules regarding future activities and future power consumption: given a previous sequence of activities \((X)\), we can infer what is the most probable next activity \((y)\), and how much energy will be consumed while performing the predicted activity, based on previous observations \((\text{Power Load}(y) \mid X)\). With this technique it is possible to extract 685 rules in total with a confidence of 100% out of our dataset. An illustrative example of the results is given in Table I. The listed 4 rules have a confidence value of 0.95, 1.0, 0.9 and 0.89, respectively.

### TABLE I: Exemplary sequence rules extracted from the households 66, 46, 62 and 67 respectively. Demonstrated are the mean and standard deviation of the power load \((PL)\) in watts per minute and the prediction horizon \((PH)\) in minutes.

<table>
<thead>
<tr>
<th>Rule</th>
<th>(\mu_{PL})</th>
<th>(\sigma_{PL})</th>
<th>(\mu_{PH})</th>
<th>(\sigma_{PH})</th>
</tr>
</thead>
<tbody>
<tr>
<td>{out, bath, media, eat, } \Rightarrow sleep</td>
<td>394.40</td>
<td>109.01</td>
<td>68.70</td>
<td>27.67</td>
</tr>
<tr>
<td>{eat, housework, eat, housework, eat} \Rightarrow relax</td>
<td>931.35</td>
<td>232.88</td>
<td>69.25</td>
<td>14.88</td>
</tr>
<tr>
<td>{sleep, eat, out, eat} \Rightarrow media</td>
<td>374.42</td>
<td>68.88</td>
<td>63.00</td>
<td>19.38</td>
</tr>
<tr>
<td>{housework, media} \Rightarrow bath</td>
<td>574.79</td>
<td>211.42</td>
<td>127.58</td>
<td>92.15</td>
</tr>
</tbody>
</table>

As shown in the above table, with the high confident sequence-based rules, we are able to infer the average power consumption of a future activity. Although the impact of the same future activities on the power consumption could differ greatly due to a high standard deviation, each activity type features a specific probability distribution of the power load. Moreover, the above 4 sequence-based rules present a predictability of the future activity in terms of the prediction horizon ranging from about 15 minutes to nearly 4 hours within the \(2\sigma\) range.

With the above rule analysis, we could gather a deep understanding of the causal factors related to activities that influence power consumption of individual households and the associations between them. These associations may be used to support end-user applications that increase the awareness of activity related power consumption. Automatically deriving causal ties between sequences of activities and power use is therefore a promising approach to improve demand side management. Furthermore, the rule-based power consumption inference is rather valuable for detecting consumption anomaly than for load forecasting in general, since the above presented rule-based power consumption inference is highly dependent on the occurrence of mined sequences and refers mainly to a discrete time interval. How to exploit mining activity sequences for the future power load will be discussed in the next section.

**IV. LOAD FORECASTING WITH ACTIVITY SEQUENCES**

In this section, we conduct a load forecasting study at household’s level, which analyzes the impact of the collected activity information and the mined frequent activity sequences on the forecasting accuracy. As mentioned in Sec. I, the activity information studied in the current research is working effectively as an important influencing factor for the load forecasting model. The question arises whether our collected activity information can bring any benefit for load forecasting and to what extent the frequent activity sequences outperform individual activities for load forecasting.
As the sequence mining study in Sec. III shown, the mined frequent activity sequences influence the power consumption of individual households to varying degrees. By means of sequence-based rules, it is possible to infer future activities and the corresponding power consumption in average, given the previous activity sequence. In the following, we demonstrate an activity-enhanced load forecasting model with SVR (Support Vector Regression), in which the mined frequent activity sequences will be investigated in comparison with the individual activities.

The collected activity samples (log entries) have very different activity durations in minutes: 97.1% are greater than 15 minutes. Considering 15-minute-interval meter reading data that are relevant for the energy markets and should be supported by the most of Advanced Metering Systems, the following load forecasting model focuses therefore on a prediction lead time of 15 minutes and can be sufficiently evaluated with 97.1% of the activity samples. This means that the frequent activity sequences must be prepared or preprocessed in such a way that they can be taken into account as an influencing input variable for the same load forecasting model.

A. Training Preparation

For the load forecasting study, we use the framework from our previous work [24]. The goal of this study is to investigate the impact of activities and activity sequences as influencing factors, rather than to compare different load forecasting algorithms. Therefore, we use only SVR through this study due to its adequate computational effort and promising forecasting performance. For the implementation, the LIBSVM library [25] is applied.

In order to analyze the capability of activity sequences as influencing factor for the load forecasting, we take activity sequence information \( S \) as input variable into account. In addition, day type \( D \) and day hour \( H \) as well as activity information \( A \) will serve as a comparison. The forecasting accuracy is evaluated for the following 3 cases of input variables: 1) \( D \) and \( H \) as input variables; 2) \( A \) as input variable; 3) \( S \) as input variable.

1) Feature Vector: The feature vector for training the forecasting model comprises not only the above input values, but also the historical consumption data. In our case, we take the past 24 hours’ consumption data for each feature vector into account, which means 96 past power consumption data points in a feature vector for the 15 minutes ahead load forecasting. Sample data of 1 month for applicable variables are used within each training process. Min-max normalization is applied to all feature elements. Day type, day hour as well as type of activity and sequence are encoded in unary into the feature vector.

- Day type information with 7 categorical values \( \in D \rightarrow 7 \) Boolean bits;
- Day hour information with 24 integer values \( \in H \rightarrow 24 \) Boolean bits;
- Activity information with 11 categorical values \( \in A \rightarrow 11 \) Boolean bits.

- Activity sequence information with \( N_i^{sup} \) categorical values \( \in S \rightarrow N_i^{sup} \) Boolean bits.

Unlike the activity types, which refer to the 11 meta activities for all households, the frequent activity sequence types are household-specific and support value dependent. Therefore, for each household with ID \( i \), \( N_i^{sup} \) bits are applied to encapsulate the household-specific activity sequences with different support values.

2) Sequence Selection: For the activity feature vector explained before, there is always at most only one Boolean bit active (set to 1), representing the activity being performed in the current time interval being analyzed. Meanwhile this is natural for the activity feature, a decision must be made when activating the bit for a sequence. For this purpose, we define the notion of an applicable sequence: a frequent sequence is applicable, when it is a subsequence of the progression of activities performed by the household in the day prior to the time interval being currently analyzed (here called the current sequence \( S_{current} \) and the current activity is contained in this frequent sequence.

When for example analyzing a time interval in which a household is performing the activity “eat”, after having performed the activities of \( S_{current} = \{\text{cooking, eat, out, relax, eat}\} \), two applicable sequences out of \( S_{current} \) are \( S_1 = \{\text{cooking, eat, out}\} \) and \( S_2 = \{\text{out, relax, eat}\} \), as both are subsequences of \( S_{current} \) and contain the current activity “eat”. When performing an unrelated activity (e.g., “bath”) both sequences would become non-applicable. A sequence which is a subsequence of \( S_{current} \), but does not contain the current activity (e.g., \( S_3 = \{\text{cooking, out}\} \)), would also not be applicable. This strategy offers a trade-off on the set of all possible frequent sequences of a household which are subsequences of \( S_{current} \) and can become active in the sequence feature vector – while accounting for the current activity being performed in the analysis. Furthermore, we mine the frequent activity sequences for all households at 5 different support levels \( \in \{0.1, 0.3, 0.5, 0.7, 0.9\} \).

B. Forecasting Tests with Activities

To study forecasting accuracy, we apply the Mean Absolute Percentage Error (MAPE) as the performance criterion. First of all, we compare the forecasting accuracy only between the following two input variables: day type \( D \) and day hour \( H \) versus activity information \( A \).

![Table II: Mean and Standard Deviation (std) of MAPE values (in %) of 21 applicable households for 15 minutes ahead load forecasting with different settings: \( D \), \( H \) and \( A \) stand for day type, day hour and activity information, respectively.](image)

Table II shows an overview of the forecasting results in terms of mean and standard deviation of MAPE values for different input settings. In this test, we used the first 2 weeks of the data set to train the SVR model, and to predict
the power consumption of the rest experiment time. Since
the 15-minute load exhibits more aperiodicity and dynamics
in comparison with the hourly or daily load, the MAPE
values for all three input settings are relatively high. However,
the highlighted MAPE values in the table indicates that the
SVR forecasting result with activity information as an input
variable outperforms the results with other input variables. The
contribution of activity information for the load forecasting has
not only been demonstrated through the mean MAPE value,
but has also been confirmed in comparison with individual
MAPE values as shown in Figure 3. On this bar chart, we
see that for all depicted individual households except for ID
73, the red bar is under the other two bars, which means
considering only activity information as input variable can
reduce the forecasting error.

C. Forecasting Tests with Sequences

The question remains whether and to what extent the
mined frequent activity sequences can improve the fore-
casting accuracy. As mentioned before, for individual house-
holds, we apply a sequence selection strategy at 5 support
levels. We then compare the forecasting results between 5
sequence variables (due to 5 varying support thresholds) and
1 activity variable.

For the following tests, we build the SVR forecasting
model with a RBF (radial basis function) kernel and train it
with sample data of 1 month. As suggested by the authors of
LIBSVM [25], we conduct a cross-validation on the 1-month
training data set to determine the best parameter $C$ and $\gamma$.

Since each household has own applicable activity se-
quenues and several households do not own any applicable
sequences at the support level 0.9, we show only the fore-
casting results of households, which exhibit rich frequent
sequence information in Figure 4. In total, more than half of the
households could not achieve load forecasting improvement
with activity sequence information as influencing variable.
Nevertheless, we did notice load forecasting enhancements
with activity sequences in comparison to activity information
alone at 3 households (ID: 56, 59 and 62) to varying degrees.
In particular, for household with ID 56, the forecasting results
with 5 different activity sequence variables all outperform the
results with individual activity information.

Figure 4 shows that the activity sequence variable at
the support level 0.7 could greatly improve the forecasting
accuracy for the household 56. So, finally, we demonstrate
the predicted load curve in comparison with the real one
for the household 56 in Figure 5. As seen in the figure, the
predicted load corresponding to sequence-driven load forecast
(in red) could converge more towards the real peak load of the
household 56 than the predicted load using individual activities
alone (in green).

To sum up, we found out that the very short-term load fore-
casting (i.e. 15 minutes ahead) at household’s level can benefit
from activity sequences to varying degrees. In other words, the
contribution of activity sequences for load forecasting accuracy
is extremely household-dependent and household-specific.

V. Conclusion

In this paper, we proposed an activity sequence driven
approach for inferring future activities and thereby enhancing
load forecasting. Using a 3-month data set consisting of
power consumption at 1-minute intervals and residents’ daily
actions in terms of activity logs, we conducted a data-driven
analysis of the relationship between the household’s power
consumption and human activity sequences. We presented a
sequence mining study to understand daily activity patterns
for individual households. Subsequently, we could show the
potential of sequence-based rules for inferring the average
power consumption of future activities. The load forecasting
analysis with activity sequence variable as an influencing factor
shown that activity sequences could improve the accuracy of 15 minutes load forecasting for individual households to varying degrees. For certain households, the activity sequence variable as influencing factor outperformed the activity variable in terms of forecasting accuracy.

As we noticed in our load forecasting study, more than half of the households could not profit from activity sequences to reduce their forecasting error. The reason is probably, insufficient data set in total for training the forecasting model on the one hand, no quality control on the activity logs on the other hand. Furthermore, we conducted a brief analysis on sequence-based rules for inferring future activities and their future power consumption. In order to develop an activity sequence enhanced load forecasting model using the rule-based power consumption inference, an in-depth study on our proposed sequence-based rule mining will be followed as our future work.

ACKNOWLEDGMENT

The part of this research are supported by Grant-in-Aid for Scientific Research (B) “Sensor Context Estimation Techniques by Semantic and Physical Layer (Principal Investigator: Sozo Inoue)”, and CREST, JST. The authors would like to thank their support.

REFERENCES


