Profiling of Household Residents' Electricity Consumption Behavior using Clustering Analysis

Christian Nordahl, Veselka Boeva, Håkan Grahn, and Marie Persson Netz

Blekinge Institute of Technology, 371 79 Karlskrona, Sweden
{christian.nordahl, veselka.boeva, hakan.grahn, marie.netz}@bth.se

Abstract. In this study we apply clustering techniques for analyzing and understanding households' electricity consumption data. The knowledge extracted by this analysis is used to create a model of normal electricity consumption behavior for each particular household. Initially, the household's electricity consumption data are partitioned into a number of clusters with similar daily electricity consumption profiles. The centroids of the generated clusters can be considered as representative signatures of a household's electricity consumption behavior. The proposed approach is evaluated by conducting a number of experiments on electricity consumption data of ten selected households. The obtained results show that the proposed approach is suitable for data organizing and understanding, and can be applied for modeling electricity consumption behavior on a household level.

Keywords: Ambient Assisted Living·Non-Intrusive Remote Monitoring

1 Introduction

The world's population is getting older. By 2050, projections state that the number of individuals over 60 will be around 2.1 billion. Keeping individuals in their own homes, and delaying their entrance to the health and elderly care systems, can help to offload costs and work from the already strained health care systems. Likewise, the elderly population often want to keep living independently at home, but they also want a sense of safety without any intrusion in their lives [13]. Remote monitoring, and assistance, is one way to provide the safety of the residents. Traditionally, remote monitoring has been performed with video surveillance. However, with the introduction of smart homes, i.e. houses with built in sensors and actuators, Ambient Assisted Living (AAL) has emerged and allows for monitoring and assistance without the use of cameras and with less intrusion of the residents' privacy.

With the adoption of smart meters in the electrical power grids, we have the opportunity to collect high resolution electricity consumption data remotely on a household level. This type of data can be used to get insight into the residents' habits and activities, with low impact and intrusion of the residents' privacy. We may detect abnormalities and changes of residents' behavior through analyzing their daily household electricity consumption. For example, dementia, and other neurodegenerative diseases, cause changes in the behavior of the individual in different ways, e.g., they can provoke insomnia, apathy, restlessness etc. [8]. We believe that changes like these, in the individual's daily behavior, can be caught by his/her electricity consumption activities.

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22750-0_78

Most current research related to household electricity consumption has mainly revolved around creating consumer profiles by clustering households together [3] and comparing different households to determine and predict abnormal consumption patterns [2]. But, there has been some research as well on household electricity consumption. For example, Zhang et al. [14] analyze energy consumption data on a household level to identify days when the residents have gone on vacation. Further, we have previously studied the use of prediction models for electricity consumption behavior [10].

In this paper, we present and evaluate a cluster analysis approach for organizing, understanding, and modeling household electricity consumption data, a continuation of our work in [9]. Our aim is to study the possibility of using the knowledge discovered by such analysis for creating consumption behavior signatures on a household level. The long-term goal is to investigate whether the created signatures can be used for identifying abnormal behavior in daily life and apply this outlier detection model in health care applications, e.g., for monitoring early stages of dementia or other neurodegenerative diseases. The developed consumption signatures can be considered as predefined activities and can be used for detecting abnormal consumption patterns in order to notify the environment (relatives and health care professionals) if early signs of dementia occurs repeatedly at home.

2 Clustering Analysis Approach

Data Pre-Processing The electricity consumption data collected in this study is gathered with a one-minute resolution and is measured in kWh (kilowatt hours). To be able to determine and profile a behavior of the household, we divide the time series into 24 hour profiles. This is a intuitive division of the data, as it allows us to capture a daily behavior which we then can analyze and use to model a routine daily behavior.

We set a maximum limit of 10% of the entire day or 20 consecutive minutes of missing data to remove that day from the data set. For the remaining profiles, we impute missing values by using linear interpolation. We then aggregate the electricity consumption data into a one hour resolution due to that resolution being more common for today's smart meters. Finally, we standardize the time series profiles using z-standardization, or Z-score, because we are more interested in the general shapes of the time series and not the actual amplitudes.

Clustering Algorithms & Validation Measures Three partitioning algorithms are commonly used for data analysis to divide the data objects into k disjoint clusters [7]: k-means, k-medians, and k-medoids clustering. The three partitioning methods differ in how the cluster center is defined. In k-means clustering, the cluster center is defined as the mean data vector averaged over all objects in the cluster. In k-medians, the median is calculated for each dimension in the data vector to create the centroid. Finally, in k-medoids clustering, which is a robust version of the k-means, the cluster center is defined as the object with the smallest sum of distances to all other objects in the cluster, i.e., the most centrally located point in a given cluster. We have used k-medoids, since having an actual consumption profile as the cluster's centroid (medoid) is more representative of the consumption behavior compared to creating a synthetic centroid.

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22750-0_78

There are two major categories in which we can divide cluster validation measure to: *external* and *internal*. External measures are used when you have prior knowledge of the data and validate according to the ground truth and internal measures validate based on the data and clusters themselves [5]. We use three internal validation measures for analyzing the data and to select the optimal clustering scheme. We have selected one validation measure for assessing compactness and separation - *Silhouette Index* [11], one for assessing connectedness - *Connectivity* [6], and one for assessing tightness and dealing with arbitrary shaped clusters - *IC-av* [1].

Distance Measures The simplest and most widely used way to measure the distance, or dissimilarity, between two data points in a n-dimensional space (in our case two time series) is to calculate the Euclidean Distance (ED). ED calculates the distance between the two time series by aligning the ith point of one time series with the ith point of the other. ED is fast but it is sensitive to outliers and it cannot identify similarities between two segments if they are shifted out of phase [4]. Therefore, we also investigate the use of Dynamic Time Warping (DTW).

DTW measures the dissimilarity of two time series in a similar way, but instead of strictly calculating point by point it allows for some elasticity. One point in one of the time series can be aligned against one or more points in the other [12], which allows the identification of similar shapes even though they are out of phase.

3 Experiments and Results

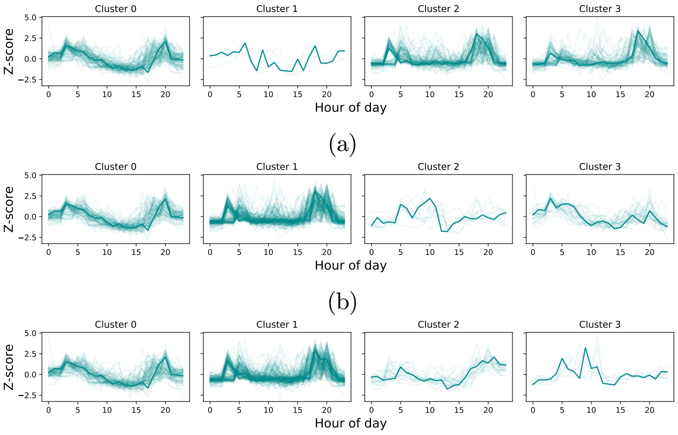
3.1 Data

We have gathered electricity consumption data from 9909 anonymous households, collected with a 1-minute interval. Initially, the household data have gone through the pre-processing stage, as explained in Section 2. Then we have selected the 10 households with the largest number of daily profiles. 3 of these 10 households contained a few clearly abnormal profiles which we excluded from further analysis. At the final stage, the 10 selected households contain between 345 and 353 daily profiles. We then selected 1 of the 10 studied households as the representative and we discuss and interpret the results obtained on its data for the rest of our study.

3.2 Results and Analysis

Estimation of the Number of Clusters We run the k-medoids clustering algorithm using the two distance metrics (ED and DTW) for all values of k between 2 and 9. The clustering algorithm is run 100 times for each k and with the cluster medoids randomly initialized. All clustering solutions are then evaluated using the cluster validation measures mentioned in Section 2. We then look upon each measure individually and in combination to determine which k is the appropriate. Based on the scores generated by the validation measures, we decide upon $k=5$ for ED and $k=4$ for DTW. However, the scores we evaluated are the best ones generated from each individual measure, i.e., the scores are not necessarily generated from the same clustering solutions.

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22750-0_78

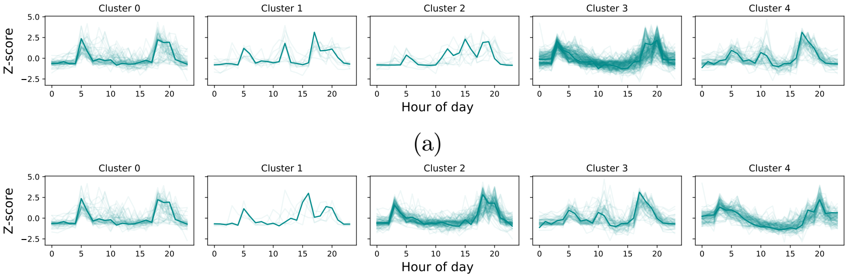


(a)

(b)

(c)

Fig. 1. Clustering solution generated by k-medoids for $k = 4$, using DTW as a dissimilarity measure and supported to be the best by; (a) IC-av, (b) SI, and (c) Connectivity.



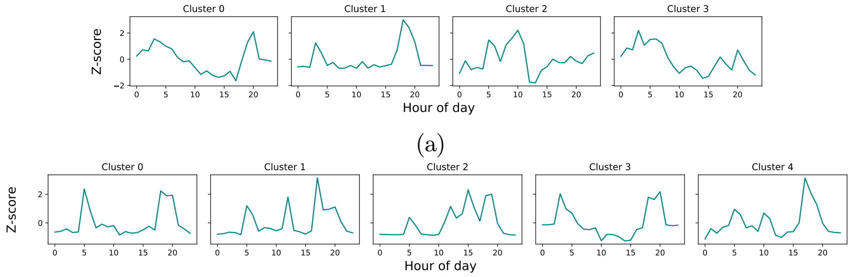
(a)

(b)

Fig. 2. Clustering solution generated by k-medoids for $k = 5$, using ED as a dissimilarity measure and supported to be the best by; (a) IC-av and Connectivity, and (b) SI.

Clustering Analysis In Figure 1 we show the clustering solutions produced by k-medoids using DTW as a distance metric. All three validation measures support different clustering solutions. Therefore, we compare the three solutions and choose one of them that will be used to analyze the produced household consumption signatures. SI and Connectivity do however, share two clusters with the same signature and both of them contain mostly the same profiles. Both these solutions have two major and two smaller clusters. IC-av, on the other hand, generates a solution with three major clusters and only one small cluster. The solutions chosen by SI and Connectivity are fairly similar, with only a few profiles difference between them. Therefore, we have chosen the clustering solution supported by SI (Figure 1(b)) as its smallest clusters contain more profiles compared to their counterparts in the solution preferred by Connectivity.

In Figure 2 we present the clustering solutions produced by ED. In this scenario, both IC-av and Connectivity support the same clustering solution, while SI has a solution of its own. However, the differences between the two solutions are not as



(a)

(b)

Fig. 3. Electricity consumption signatures created from the cluster medoids from the chosen clustering solutions generated by k-medoids with (a) $k=4$ (DTW) and (b) $k=5$ (ED).

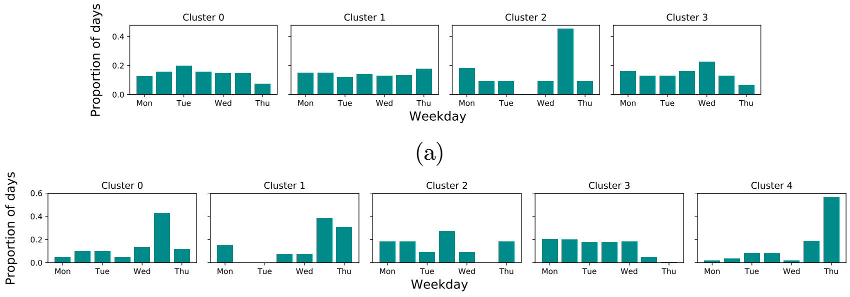
distinct as in the case of DTW. We notice that they share the same signature for two of the clusters, and one additional cluster has a very similar signature. The solution promoted by SI does not have an equally as large cluster as the other solution has. In addition, in the solution proposed by IC-av and Connectivity we have two small clusters. In case of ED, we use the majority rule for choosing which clustering solution is the best and thereby it is the one proposed by IC-av and Connectivity (Figure 2(a)).

Consumption Behavior Signatures The produced cluster centroids, which can be seen as the signatures of the electricity consumption habits of the household, are shown in Figure 3. We can see that the two different distance measures support different consumption signatures. For instance, DTW (Figure 3(a)) focuses more on the general shape of the electricity consumption profiles, while ED (Figure 3(b)) favours more the exact time of the days when the consumption peaks are happening. This supports our expectations, since DTW is an elastic measure that stretches the time series in the time axis to find an optimal alignment. Evidently, the different distance measures favour different electricity consumption profiles and logically, this will affect the intended analysis and built signatures. For example, we notice clear morning and evening consumption peaks recognized by Cluster 1 of the DTW solution (see Figure 3(a)) and Clusters 0 and 3 of the ED signatures (see Figure 3(b)), respectively. However, the additional consumption peak in the middle of the day seen in Cluster 1 of the ED solution is not clearly presented in any of the four signatures supported by DTW.

Context Based Signatures In order to investigate further the electricity consumption behavior of households, we create context based signatures that represent how the weekdays are distributed between each cluster. These signatures are shown in Figure 4. They can be used to further improve and refine the consumption behavior model and allow us to gain additional knowledge about the household's behavior.

Comparing the distribution of weekdays produced by DTW (see Figure 4(a)) and ED (see Figure 4(b)), it is apparent that DTW divides the days more evenly than ED. The signatures of the three larger clusters, 0, 1, and 3, are fairly monotonic. ED, on the other hand, has a more distinct separation between working days and weekends.

DOI: 10.1007/978-3-030-22750-0_78



(a)

(b)

Fig. 4. Context based signature of the weekdays from the clustering solution generated generated by k-medoids with (a) $k=4$ (DTW) and (b) $k=5$ (ED).

Clusters 0, 1, and 4 all contain more weekend days, while Cluster 3 has almost only working days. Cluster 2, which is the smallest cluster only containing 11 signatures, has a more diverse spread of its days. It is further interesting to notice that the signatures of Clusters 0 and 3 (see Figure 3(b)) are very similar, which can be an indication that these could be merged into a single cluster. However, this is not strongly supported by the context presented in Figure 4(b), since the first cluster present a typical working day consumption behavior while the second one is more representative for the weekends.

3.3 Discussion

The produced electricity consumption signatures are representatives of the *current* electricity consumption behavior of the residents. To detect changes in the residents' behavior, we can apply our approach on a new portion of data presenting the electricity consumption for the next time period. If new signatures are created through the clustering process, this might be a sign of a new behavior of the resident.

The proposed method can also be used to produce better prediction models. The generated clusters give a clearer distinction between normality and abnormality of the electricity consumption profiles. For example, in Figure 2(a) it is clear that clusters 0, 3, and 4, contain a majority of the electricity consumption profiles. Training the prediction models only on the contents of these clusters would give a more accurate model.

As mentioned before, it is beneficial to use some elasticity when comparing the electricity consumption profiles. Using DTW, we allow for changes in time for the individual electricity consumption profiles. However, we may cluster some signatures which probably should not be regarded as similar, e.g., if a resident is sick and stays in bed for a few extra hours and then performs his/her daily routines. We would like to identify this as a behavioral change, but DTW identifies this as normal. Introducing a time window for DTW to warp would remedy this.

4 Conclusions and Future Work

In this paper we propose a clustering analysis approach for profiling a households electricity consumption behavior. The proposed approach is evaluated on real electricity consumption data from 10 anonymous households. The results show that we can create electricity consumption signatures that model electricity consumption behaviors of the household residents. Further, we have found that Euclidean distance (ED) produce clusters that are more focused on the exact times of consumption peaks, Dynamic Time Warping (DTW) was better at identifying the shapes of the consumption peaks. We have also identified that ED has a clear distinction between working days and weekend days between the clusters, while DTW has a more monotonic distribution of days.

We are currently in the final stages of collecting both electricity and water consumption from a set of elderly residents with the help of our local elderly care system. The subjects will be continuously interviewed and monitored during the study to be able to accurately label the data, i.e. changes in their behavior. The collected data will be used for further evaluation and validation of the approach proposed in this study.

References

1. Baya, A.E., Granitto, P.M.: How many clusters: A validation index for arbitrary-shaped clusters. IEEE/ACM Trans. on Comput. Biol. and Bioinformatics **10**(2), 401–414 (2013)
2. Chalmers, C., Hurst, W., Mackay, M., Fergus, P.: Profiling users in the smart grid. In: The Seventh International Conference on Emerging Networks and Systems Intelligence (2015)
3. Chen, T., Mutanen, A., Järventausta, P., Koivisto, H.: Change detection of electric customer behavior based on AMR measurements. In: PowerTech, 2015 IEEE Eindhoven. pp. 1–6. IEEE (2015)
4. Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., Keogh, E.: Querying and mining of time series data: experimental comparison of representations and distance measures. Proceedings of the VLDB Endowment **1**(2), 1542–1552 (2008)
5. Halkidi, M., Batistakis, Y., Vazirgiannis, M.: On clustering validation techniques. Journal of intelligent information systems **17**(2-3), 107–145 (2001)
6. Handl, J., Knowles, J., Kell, D.B.: Computational cluster validation in post-genomic data analysis. Bioinformatics **21**(15), 3201–3212 (2005)
7. MacQueen, J., et al.: Some methods for classification and analysis of multivariate observations. In: 5th Berkeley Symp. on mathematical statistics and probability. vol. 1, pp. 281–297 (1967)
8. Mega, M.S., Cummings, J.L., Fiorello, T., Gornbein, J.: The spectrum of behavioral changes in alzheimer's disease. Neurology **46**(1), 130–135 (1996)
9. Nordahl, C., Boeva, V., Grahn, H., Netz, M.: Organizing, visualizing and understanding households electricity consumption data through clustering analysis. In: 2nd workshop on Aging, Rehabilitation and Independent Assisted Living, IJCAI Workshop (2018)
10. Nordahl, C., Persson, M., Grahn, H.: Detection of residents' abnormal behaviour by analysing energy consumption of individual households. In: Data Mining Workshops (ICDMW), 2017 IEEE International Conference on. pp. 729–738. IEEE (2017)
11. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics **20**, 53–65 (1987)
12. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE trans. on acoustics, speech, and signal processing **26**(1), 43–49 (1978)
13. Zagler, W.L., Panek, P., Rauhala, M.: Ambient assisted living systems-the conflicts between technology, acceptance, ethics and privacy. In: Dagstuhl Seminar Proceedings. Schloss Dagstuhl-Leibniz-Zentrum fr Informatik (2008)
14. Zhang, Y., Chen, W., Black, J.: Anomaly detection in premise energy consumption data. In: Power and energy society general meeting, 2011 ieee. pp. 1–8. IEEE (2011)

DOI: 10.1007/978-3-030-22750-0_78