Available online at www.sciencedirect.com

ScienceDirect

Energy Procedia 88 (2016) 754 – 760

Energy Procedia

CUE2015-Applied Energy Symposium and Summit 2015: Low carbon cities and urban energy systems

K-means based cluster analysis of residential smart meter measurements

Ali Al-Wakeel[a,*], Jianzhong Wu[a]

[a] *School of Engineering, Cardiff University, Cardiff CF24 3AA, United Kingdom*

Abstract

A clustering module based on the *k*-means cluster analysis method was developed. Smart meter based residential load profiles were used to validate the clustering module. Several case studies were implemented using daily and segmented load profiles of individual and aggregated smart meters. Simulation results defined in terms of the relationship between the clustering ratio and the segmentation time window reveal that the minimum clustering ratio is obtained for the shortest time window of segmentation. Results also show that a small number of clusters is recommended for highly correlated load profiles.

Peer-review under responsibility of the organizing committee of CUE 2015

Keywords: Smart meter measurements; *k*-means; cluster analysis

1. Introduction

Smart meter measurements collected at 15-min, 30-min or 1-hour time intervals increase the available amount of data that describe power consumption of residential, small commercial and small industrial customers. The transformation of smart metering data into practical information greatly improves the operation, planning and control of distribution networks. Statistical, engineering, time-series, and cluster analysis methods are used to extract such important information as load profile classes [1] from the measurements of smart meters.

The application of the *k*-means cluster analysis method to group load profiles of residential customers was reported in literature [1]–[7]. The target of the previous research was to cluster daily load profiles. However, this paper investigates the clustering of daily and segmented load profiles of individual and aggregated residential customers. Segmented load profiles are profiles that have a time span less than or

* Corresponding author. Tel.: +44-2920-870670.
E-mail address: al-wakeelas@cardiff.ac.uk.

Peer-review under responsibility of the organizing committee of CUE 2015
doi:10.1016/j.egypro.2016.06.066

equal to 24 hours. The outputs of the proposed k-means based clustering module can be used for load estimation where missing and future smart meter measurements are estimated; or for the improvement of the Time of Use (ToU) tariff design. In this paper, the proposed k-means based clustering module is presented, and simulation results are analyzed. The optimal time window of segmenting load profiles is defined based on a comprehensive analysis of simulation results.

2. Cluster Analysis Methods

Clustering is defined as the grouping of similar objects. A given set of load profiles is grouped into a number of clusters such that profiles within the same cluster are similar to each other. At the same time, load profiles that are assigned to different clusters are as dissimilar as possible. Clustering implies that the number of output clusters is less than or equal to the number of input load profiles.

A large number of cluster analysis methods were developed as a result of the wide range of the existing applications of clustering. Applications of cluster analysis methods include data mining, pattern recognition, and clustering based estimation. Cluster analysis methods are broadly categorized into hierarchical and partitional clustering methods. Hierarchical methods [8] group a given dataset of load profiles into the required number of clusters through a series of nested partitions. This results in a hierarchy of partitions leading to the final cluster(s).

Partitional methods on the other hand aim to group load profiles into a number of clusters by optimizing an objective function. The intra-cluster sum of squared distances is the objective function that is minimized. Partitional clustering imposes that the required number of clusters must be predefined or known in advance. Partitional cluster analysis methods are called center-based methods because each cluster is represented by a corresponding center. The center of a cluster is often seen as a summary description of all load profiles contained within that specific cluster. Partitional methods are very efficient for clustering large and high-dimensional datasets. As a consequence, partitional methods are preferred for clustering daily load profiles of residential customers [9], [10].

3. Proposed Cluster Analysis Module

The k-means method [12] is one of the most used partitional cluster analysis methods. This method is an iterative process that groups n load profiles – each comprised of T half-hourly measurements – into k clusters, by minimizing the intra-cluster sum of squares demonstrated in equation (1).

$$J = \sum_{j=1}^{k} \sum_{i=1, i \in j}^{n} \left\| \mathbf{LP_i} - \mathbf{CC}_j \right\|^2 \quad (1)$$

$\mathbf{LP_i}$ is the i^{th} load profile, $i = 1, 2, 3, \dots, n$, and $\mathbf{CC_j}$ is the j^{th} cluster center, $j = 1, 2, 3, \dots, k$. The i^{th} load profile is described as $\mathbf{LP_i} = \{lp_i(t), t = 1, 2, 3, \dots, T\}$. Similarly, the j^{th} cluster center is defined as $\mathbf{CC_j} = \{cc_j(t), t = 1, 2, 3, \dots, T\}$.

The inputs of the proposed k-means based clustering module include load profiles of residential customers and the maximum number of clusters. The maximum number of clusters is always equal to the number of input load profiles. In the case that the maximum number of clusters is reached, each load profile will have its own cluster. Load profiles are the respective centers of their clusters in this case. At each iteration of the k-means, the average Euclidean distance is calculated between the load profiles and the cluster centers according to equation (2). This results in the assignment of each load profile to the cluster that has the nearest center.

$$d\left(\mathbf{LP_i}, \mathbf{CC_j}\right) = \sqrt{\frac{\sum_{t=1}^{T} \left(lp_i(t) - cc_j(t)\right)^2}{T}} \quad (2)$$

The mean value of the root-mean-square errors (RMSE) between load profiles and their corresponding cluster centers is used as a criterion to determine the required number of clusters. The number of clusters

is iteratively incremented until the mean RMSE falls below an error threshold. The error threshold is defined in equation (3).

$$Error\ threshold = \frac{X \times Average\ consumption\ of\ input\ profile}{100} \quad (3)$$

$X = 1, 2, 3, \ldots, 10$. The outputs of the clustering module are represented by the number of clusters, the cluster centers, and the assignment of load profiles to their respective clusters. A cluster center is determined in terms of the average values of all load profiles assigned to this specific cluster, calculated at each half-hourly time step.

4. Clustering Methodology

Pycluster [12], an open source cluster analysis software was used to develop the clustering module in Python 2.7. Residential load profiles based on actual smart meter measurements were used to assess the performance of the proposed module. The load profiles were obtained from the Irish Smart Metering [13] Customer Behavior Trials (CBT). Daily load profiles of each smart meter comprise of 48 half-hourly measurements. The first measurement collected at 12:30am represents the average active power consumption between 12:00am and 12:30am. The last measurement taken at 12:00am comprises the average value of the active power consumed between 11:30pm and 12:00am.

Load profiles of 100 residential smart meters collected over the period extending from 20 July until 9 August 2009 were used in the present study. These were divided into training period and test period profiles.

4.1. Training Period Profiles

Load profiles collected over the period between 20 and 26 July 2009 were used to train the proposed k-means based clustering module. These profiles were applied directly as inputs to the clustering module. As a result, an output comprising a number of clusters accompanied with their corresponding centers was obtained.

4.2. Test Period Profiles

Ten different sub-periods each with three consecutive days cover the duration of the test period. The test period extends between 27 July and 9 August 2009. A load profile of the test period was allocated (i.e., re-clustered) to the nearest center obtained from the clustering of training period profiles. The allocation was based on the minimum average Euclidean distance between the test period profile and the cluster centers of the training period. The re-clustering error calculated between the test profiles and their respective training cluster centers was quantified in terms of the maximum absolute error (AE) and the RMSE.

5. Simulation Results

Cluster centers were acquired through the clustering of load profiles of the training period. Load profiles of the test period were allocated to the nearest cluster center. The test period profiles and training period centers had the same length. Daily and segmented profiles of individual and aggregated smart meters were separately clustered and then re-clustered. As compared to a daily load profile that consists of 48 half-hourly measurements, even time windows in the range of [2, 24] hours were used to create the segmented load profiles of the training period and the test period. The approach of producing segmented profiles is illustrated in Figure 1.

The number of segments was calculated according to equation (4)

$$number\ of\ segments = (n \times T) - 2r + 1 \quad (4)$$

	HH01	HH02	HH03	HH04	HH05	HH06	⋮	HH43	HH44	HH45	HH46	HH47	HH48
Profile 1													
Profile 2													
Profile 3													
Profile 4													
Profile 5													
Profile 6													
⋮							⋯						
Profile XX-5													
Profile XX-4													
Profile XX-3													
Profile XX-2													
Profile XX-1													
Profile XX													

Fig. 1 Segmentation of daily load profiles

given that r is the time window (in hours). In this manner, load profiles of the training period can be described as either seven daily load profiles, or a set of 333-segmented profiles each with a 2-hour span, for instance.

5.1. Clustering of Daily Load Profiles

Results of clustering the daily load profiles of individual smart meters showed that the clustering ratio was not impacted by the changes in error threshold. The clustering ratio that is defined as the ratio between the number of output clusters to the number of input profiles attained approximately the same range of values despite increasing the error threshold from 1% to 10% of the average consumption. These results were obtained for the majority of individual smart meters.

The power demand of individual residential customers is extremely dynamic and unpredictable. A high degree of correlation – accompanied with minimum correlation error – between different daily consumption patterns of the same customer is hard to establish. Therefore, each daily load profile was assigned to its own cluster, resulting in clustering ratio of unity.

Clustering the daily load profiles of aggregated smart meters and a small count of individual smart meters revealed a significant reduction in the clustering ratio when the error threshold was increased from 1% to 10%. This can be clearly observed in Figure 2. The figure depicts the relationship between the clustering ratio and the clustering error threshold when training load profiles of aggregated smart meters were clustered.

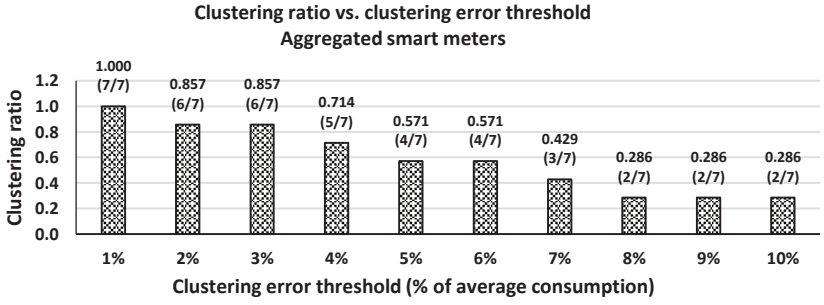

Fig. 2 Clustering ratio vs. clustering error threshold – aggregated smart meters

The numbers in parenthesis describe the number of output clusters followed by the number of input profiles. Smart meters that have highly correlated consumption patterns were found to have small numbers

of clusters when high values of clustering error thresholds were used. For different clustering error thresholds, a clustering ratio of unity was attained and therefore the same cluster centers were obtained. Consequently, the errors of re-clustering the test daily profiles of individual smart meters remain unaltered. As compared to their values at small clustering error thresholds, the errors of re-clustering the test profiles of aggregated smart meters are found to decrease at large clustering error thresholds. These results can be interpreted in terms of that the load profiles of the test period and the training cluster centers obtained for large values of the clustering error threshold were highly correlated. A small number of clusters was obtained for large error thresholds of the training period. The number of profiles per cluster in this case was greater than the number of profiles per cluster for small clustering error thresholds. This implies that clusters with a large number of profiles will retain centers that exhibit more information about the consumption trends than the centers of clusters that encompass one or two profiles.

The relation between RMS re-clustering errors of aggregated profiles for the different combinations of test periods is depicted in Figure 3. The vertical striped bar represents the re-clustering error when a clustering error threshold of 1% of the average consumption was used to obtain the cluster centers, while the horizontal striped bar refers to the case of cluster centers obtained when an error threshold of 10% of the average consumption was applied.

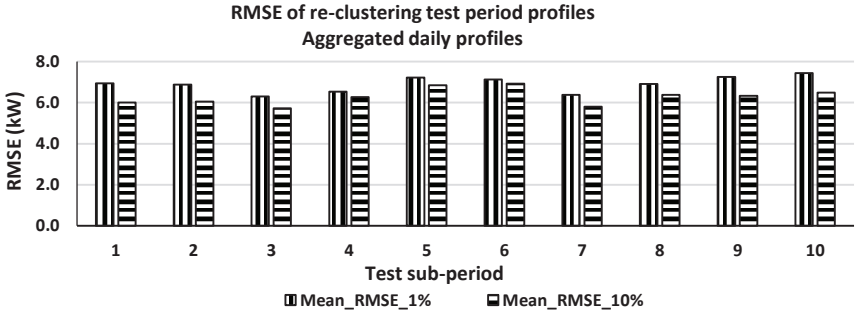

Fig. 3 RMS re-clustering error of test profiles - aggregated smart meters

5.2. Clustering of Segmented Load Profiles

Segmented profiles of individual and aggregated smart meters were used to train the clustering module and hence obtain the required cluster centers. Clustering results show that the minimum value of the clustering ratio was observed at the 2-hour time window for individual and aggregated smart meters. This means that the minimum number of clusters is attained at the aforementioned segmentation time window. Maximum values of the clustering ratio were observed when the length of the segmented load profiles of individual and aggregated smart meters was in the range of [18, 24] hours.

The clustering ratios for different segmentation time windows are illustrated in Figure 4. The results depicted in Figure 4 represent the clustering ratio attained when test profiles of aggregated smart meters were clustered using an error threshold of 10% of average consumption. Figure 5 shows a box-whisker plot of the maximum AE of re-clustering the segmented test profiles of the aggregated smart meters. It is clearly shown in Figure 5 that minimum values of maximum absolute re-clustering error were obtained at the 2-hour time window. An illustration of the RMSE of re-clustering segmented test profiles of aggregated smart meters is shown in Figure 6. In terms of RMS re-clustering error distribution, Figure 6 unveils that the minimum values were also acquired at the 2-hour segmentation time window. The dark shaded boxes in Figure 5 and Figure 6 represent the re-clustering errors that lie between the first quartile and the median of

the re-clustering errors of the test profiles of aggregated smart meters. The first quartile is defined as the middle value of error between the minimum and the median values of the re-clustering errors.

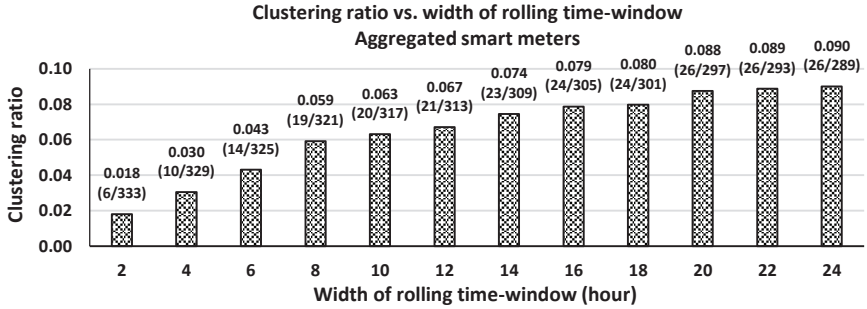

Fig. 4 Clustering ratio for different time windows – aggregated smart meters

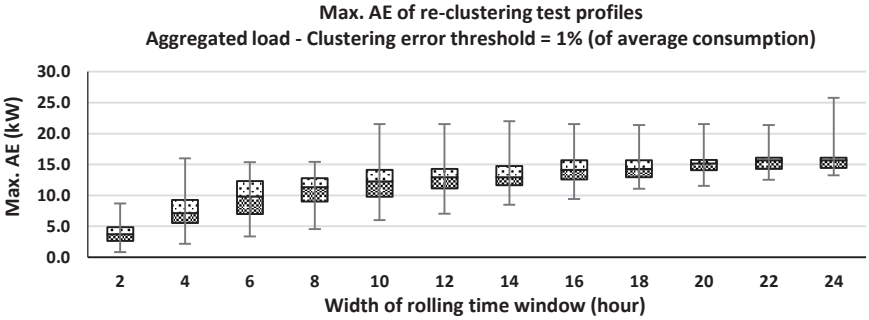

Fig. 5 Error distribution of maximum AE - aggregated smart meters

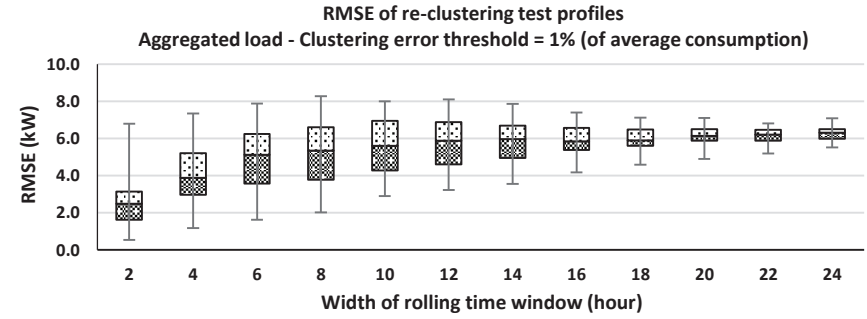

Fig. 6 Error distribution of RMSE - aggregated smart meters

The light shaded boxes correspond to the re-clustering errors whose values are in the range bounded by the median and the third quartile of the re-clustering error. The third quartile is defined as the mid-error between the median and the maximum value of the re-clustering error [14]. The dark-shaded and light-shaded box are separated by the median value of the re-clustering error.

6. Conclusions

A clustering module based on k-means cluster analysis method was developed. Detailed of the clustering results reveals the advantage to have small numbers of clusters when the high correlation between the clustered daily profiles is observed. For extremely dynamic residential load profiles, it is preferred to have large numbers of clusters since these will reduce there-clustering errors. Results of re-clustering segmented load profiles show significant correlation between segmentation time window and minimum values of re-clustering error. Minimum values of the re-clustering error were obtained at the shortest time window of segmentation.

Acknowledgements

The authors gratefully acknowledge the P2P-smarTest Programme for the partial financial support of this work through the European Commission HORIZON 2020 grant.

References

[1] F. McLoughlin, A. Duffy, and M. Conlon, "A clustering approach to domestic electricity load profile characterisation using smart metering data," *Appl. Energy*, vol. 141, pp. 190–199, 2015.
[2] M. N. Q. Macedo, J. J. M. Galo, L. a. L. Almeida, and A. C. C. Lima, "Typification of load curves for DSM in Brazil for a smart grid environment," *Int. J. Electr. Power Energy Syst.*, vol. 67, pp. 216–221, 2015.
[3] I. Benítez, A. Quijano, J. L. Díez, and I. Delgado, "Dynamic clustering segmentation applied to load profiles of energy consumption from Spanish customers," *Int. J. Electr. Power Energy Syst.*, vol. 55, pp. 437–448, 2014.
[4] J. D. Rhodes, W. J. Cole, C. R. Upshaw, T. F. Edgar, and M. E. Webber, "Clustering analysis of residential electricity demand profiles," *Appl. Energy*, vol. 135, pp. 461–471, 2014.
[5] M. Espinoza, C. Joye, R. Belmans, and B. De Moor, "Short-Term Load Forecasting , Profile Identification , and Customer Segmentation : A Methodology Based on Periodic Time Series," *IEEE Trans. Power Syst.*, vol. 20, no. 3, pp. 1622–1630, 2005.
[6] G. J. Tsekouras, N. D. Hatziargyriou, and E. N. Dialynas, "Two-Stage Pattern Recognition of Load Curves for Classification of Electricity Customers," *IEEE Trans. Power Syst.*, vol. 22, no. 3, pp. 1120–1128, 2007.
[7] T. Räsänen, D. Voukantsis, H. Niska, K. Karatzas, and M. Kolehmainen, "Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data," *Appl. Energy*, vol. 87, no. 11, pp. 3538–3545, 2010.
[8] R. Mena, M. Hennebel, Y.-F. Li, and E. Zio, "Self-adaptable hierarchical clustering analysis and differential evolution for optimal integration of renewable distributed generation," *Appl. Energy*, vol. 133, pp. 388–402, 2014.
[9] S. Bandyopadhyay and S. Saha, *Unsupervised Classification: Similarity Measures, Classical and Metaheuristic Approaches, and Applications*. Springer-Verlag, 2013.
[10] G. Gan, C. Ma, and J. Wu, *Data Clustering: Theory, Algorithms, and Applications*. American Statistical Association (ASA) & Society for Industrial and Applied Mathematics (SIAM), 2007.
[11] D. Hsu, "Comparison of integrated clustering methods for accurate and stable prediction of building energy consumption data," *Appl. Energy*, vol. 160, pp. 153–163, 2015.
[12] M. J. L. de Hoon, S. Imoto, J. Nolan, and S. Miyano, "Open source clustering software," *Bioinformatics*, vol. 20, no. 9, pp. 1453–1454, 2004.
[13] Commission for Energy Regulation, "Electricity Smart Metering Customer Behaviour Trials (CBT) Findings Report - CER11080a," Dublin, 2011.
[14] G. Der and B. S. Everitt, *Basic Statistics Using SAS® Enterprise Guide*, 1st ed. SAS Publishing, 2007.

Biography

Ali received his BSc and MSc from Baghdad University in 2005 and 2008 resepectively. Currently, he is a final year PhD student in the School of Engieering – Cardiff University. His research interests include Smart Grids and Multi Vector Energy Distribution Systems.