Multi-Layered Clustering for Power Consumption Profiling in Smart Grids

Omar Y. Al-Jarrah, Yousof Al-Hammadi, Paul D. Yoo, and Sami Muhaidat

***Abstract*—Smart Grids (SGs) have many advantages over traditional power grids as they enhance the way electricity is generated, distributed, and consumed by adopting advanced sensing, communication and control functionalities that depend on power consumption profiles of consumers. Clustering algorithms (*e.g.*, centralized clustering) are used for profiling individual's power consumption. Due to the distributed nature and ever growing size of SGs, it is predicted that massive amounts of data will be created. However, conventional clustering algorithms neither efficient enough nor scalable enough to deal with such amount of data. In addition, the cost for transferring and analyzing large amounts of data is expensive high both computationally and communicationally. This paper thus proposes a power consumption profiling model based on two levels of clustering. At the first level, local power consumption profiles are derived, which are then used by the second level in order to create a global power consumption profile. The followed approach reduces the communication and computation complexity of the proposed two level model and improves the privacy of consumers. We point out that having good knowledge of the local power profiles leads to more effective prediction model and cost-effective power pricing scheme, especially in a heterogeneous grid topology. In addition, the correlations between the local and global profiles can be used to localize/identify power consumption outliers. Simulation results illustrate that the proposed model is effective in reducing the computational complexity without much affecting its accuracy. The reduction in computational complexity is about 52% and the reduction in the communicational complexity is about 95% when compared to the centralized clustering approach.**

***Index Terms*—Smart grid, power consumption profiling, clustering.**

I. Introduction

FOR many years, conventional power grids have been used to provide electricity to consumers. However, with the growing demands of electricity, along with the diminishing fossil fuels and the environmental effect (*e.g.*, GreenHouse Gas (GHG) emissions) related to electricity generation, developing more efficient, reliable, and sustainable power grids has become a necessary need [1]. The phenomenal advances that continue to be made in the various facets of Information and Communication Technology (ICT) (*e.g.*, Wireless Sensor Networks (WSNs), Internet of Things (IoT)) enable the development of the next generation of power grids, namely Smart Grids (SGs).

O. Y. Al-Jarrah, Y. Al-Hammadi, and S. Muhaidat are with the Department of Electrical and Computer Engineering, Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates, e-mail: {omar.aljarrah, yousof.alhammadi, sami.muhaidat}@kustar.ac.ae.

P. D. Yoo is with Centre for Electronic Warfare, Information and Cyber (EWIC), Cranfield University, Defence Academy of the United Kingdom, Shrivenham, Swindon, SN6 8LA, United Kingdom, email: p.yoo@cranfield.ac.uk.

SGs use a two-way flow of power and data between devices (*e.g.*, substations, transformers, and switches) connected to the grids, in order to automate and facilitate the power flow optimization in terms of economic efficiency, reliability and sustainability [2]. To this end, each consumer location needs to be equipped with a smart meter for monitoring, measuring, and communicating the bi-directional flow of power on request or on schedule. A Supervisory Control And Data Acquisition (SCADA) system controls the grid operation by adjusting and controlling each device connected to the grid.

Although SGs introduce many advantages over conventional power grids, their utility depends heavily on the data gathered from the devices (*e.g.*, smart meters) connected to them. Data of smart meters contains correlations, trends, and patterns that are important for power consumption management, as well as the stability of grids [3]. For example, it is possible to predict the peak power usage based on the power consumption profiles of consumers and data of smart meters, enabling the supplier to address the grid power demands.

Power consumption patterns of different types of consumers vary based on the type of the consumer (*e.g.*, commercial, industrial, domestic). However, even the power consumption pattern of the same type of consumers might be different [4]. To address this, power consumption profiling, which refers to a power consumption for a consumer over a period of time, is performed [5]. This enables producing, planning, and provisioning of personalized power services based on the knowledge of the consumers' power consumption profiles [4], [6].

Clustering is the core technique of consumers power consumption profiling in SGs [5]. The main idea is to partition power consumption patterns into groups so that patterns in the same group are more similar to each other than patterns in other groups [7].

Several clustering methods have been explored in the context of consumer power profiling, such as K-means [8]–[10], Fuzzy c-means [11], hierarchical clustering method [12], and others [7]. Such methods require that all data to be located at a central site where they are analyzed. However, this approach (*i.e.*, centralized clustering) cannot be applied in the case of multiple distributed datasets, unless all data are transferred to a single location and then clustered [13]. In addition, the centralized clustering approach is costly and energy-inefficient because it: i) increases the amount of data that need to be transfered to a centralized processor, ii) requires investment in computing systems with high memory capabilities, and iii) potentially increases the number of computations.

Due to the ever growing SGs and their distributed nature, huge amounts of sensory data (*i.e.*, big data) is expected to

be produced and collected. However, as the size of the data increases, the corresponding computational cost increases as well. Therefore, not only the quality of clustering is important, but also the corresponding efficiency and scalability of the consumer power profiling model is important.

In this paper, we propose a multi-layered clustering model for SG applications. The proposed model consists of two levels of data clustering. Using the proposed model, the power consumption data of each consumer is profiled both locally, as part of a smaller scale grid (*e.g.*, microgrid, neighborhood grid), and globally, as part of the whole SG. The overall complexity is reduced by only using the representative local power profiles of the small-scale grids in the second level of clustering (*i.e.*, global clustering). In addition, since only the local power consumption profiles are transfered to the central processor of the smart grid, the privacy of the consumers is enhanced.

The rest of this paper is organized as follows: Section II provides a background of SGs and surveys some of related studies of power consumption profiling. Section III introduces the proposed multi-layered clustering model and provides a complexity analysis of the proposed model. Section IV describes the experiments and presents the results. Section V concludes and discusses future work.

II. Background

SG introduces a number of new technologies, concepts, and ideas that improve reliability and reduce costs related to power production and distribution [14]. The SG enables the integration of distributed renewable energy generation, storage equipment, and massive utilization of electric vehicles, which poses further challenges on efficient operation of the electricity grid [15]. In addition, it facilitates the users' participation to the optimization of the power consumption, via Demand Response (DR) algorithms [16].

Under the umbrella of DR, several pricing schemes have been proposed. The general idea of these schemes is to encourage users to shift their usage of high-power appliances to off-peak hours by providing economic incentives [17]. The variable pricing schemes of electricity can be either based on historical demand data or real-time demand. For instance, real-time pricing is based on the real-time market price of electricity. On the the other hand, Time of Use (ToU) is based on seasonal and daily demand data. In both cases, the price of electricity during high demand time is higher than the price of electricity during low demand time [1]. For the effective energy pricing in heterogeneous grid topologies, the concept of locational marginal prices has been introduced. It enables the determination of different prices for different areas of the electricity grid, which depends on parameters such as the line capacities and type of local loads [18], [19]. In all cases, without the installation of smart meters and the bi-directional data flow in smart grid, the deployment of different pricing schemes is not feasible. The network of smart meters is known as the Advanced Metering Infrastructure (AMI) [1]. Fig.1 shows a typical AMI structure.

Based on the data collected from smart meters, it is possible to predict power load based on power consumption patterns.

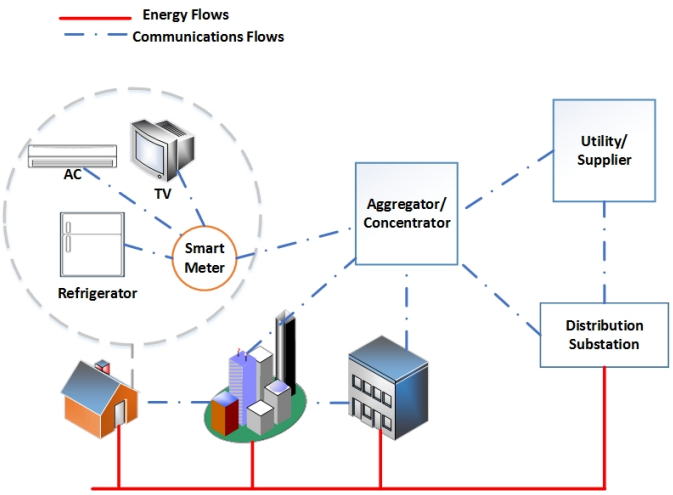


Fig. 1. A typical AMI structure: The smart meter collects the power consumption of the electricity appliances and sends the control commands to them if necessary. The data collected by the smart meters in different buildings is transmitted to a data aggregator. This aggregator could be an access point or gateway. This data can be further routed to the electric utility or the distribution substation [20].

By knowing the consumption patterns, consumers' demands might be shifted to less expensive time-slots so as to reduce the energy expenses of the consumer and to reduce the peak-to-average ratio of the grid [1].

Clustering is the core technique of power consumption profiling in SGs [5]. Several studies in the literature have discussed the application of clustering algorithms in power consumption profiling. It has been shown in [8] that clustering of consumers' power consumption patterns can be used to improve load forecasting accuracy. Chicco, *et al*. [21] tested the performance of five frequently applied clustering algorithms (K-means, fuzzy K-means, hierarchical clustering, modified follow-the-leader and SOM) for consumers power profiling. The results show the superiority of the modified follow-the-leader and hierarchical clustering as they can handle isolated uncommon power patterns. Mutaneen, *et al*. [22] proposed an Iterative Self-Organizing Data-Analysis Technique Algorithm (ISODATA) for load profiling by considering the dependency of the load patterns on temperature. Piano, *et al*. [23] used several subspace projection methods to capture subspaces of load diagrams and get in the load profiling. Tsekouras, *et al*. [24] developed a two-stage pattern-recognition methodology for the classification of electricity customers.

Although centralized clustering approaches have been used for power consumption profiling in SGs, they require that data to be processed at a single sit (central processor at the utility/supplier premises), which increases the communication overhead and the computational complexity of the clustering process, dramatically. This is because the complexity of the clustering process is a function of the number of datapoints. As the number of datapoints increases, the corresponding complexity increases as well. In contrast, distributed data-processing, fits well the concept of SG where the computations of the system could be distributed among the devices connected to the grid. Rodrigues and Gama. [25] proposed a

distributed clustering in SG. However, their approach requires residential units' participation in the clustering process. They assume that smart meters have computational capabilities to perform data clustering. However, this approach is not preferable as it increases the communication overhead since data should be communicated with all participated devices. In addition, it poses further security challenge as the consumer's data is vulnerable to cyber-attacks (*e.g.*, hijacking, eavesdropping) while communicating it with other devices in the grid.

Although the computational complexity and the scalability of power profiling model are of great interest when dealing with large-scale data, most of the existing studies in the application of clustering algorithm in SGs focus mainly on developing a more accurate centralized power profiling model. On the other hand, the proposed model in this paper aims to reduce the computational complexity of the power profiling model, while maintaining comparable performance. In addition, the proposed model mitigates the previously mentioned disadvantages of centralized data clustering by:

1) processing the collected data locally at the aggregation level, which gives a better prediction model and reduces the overall computation and communication costs;
2) enhancing the privacy of the consumers, only the local power profiles are sent to the central processor.

III. Proposed Model

This section introduces the proposed multi-layered clustering model. Then, it provides a complexity analysis of the proposed model.

A. Multi-Layered Clustering Model

Consumers' power consumption profiling, which adopts data clustering, plays a pivotal role in SGs as it is used for load forecasting, bad data correction, determination of the optimal energy resources scheduling, and power pricing [5]. Clustering algorithms discover patterns among the data based on a similarity criterion/measure. For instance, the K-means clustering algorithm partitions the data into mutually exclusive clusters of similar datapoints, aiming to maximize the intra-cluster similarity and minimize the inter-clusters similarity.

By knowing the derived profiles, normal consumers' behaviors can be recognized, which allow consumers and power suppliers to agree on consumption strategies that are more economically beneficial for both of them [6].

The proposed model consists of two levels of data clustering. The first level aims to make sense of data locality by finding representative patterns and local power profiles at data aggregation level (*i.e.*, aggregator). In this context, we define a layer l as an instance of K-means algorithm that process data of a certain region and collected by an aggregator l. The collected power consumption data at each aggregator is clustered into mutually exclusive clusters where a pattern belongs to only one cluster. Based on the observation that if patterns are located in the same cluster, they are likely to belong to the same type or have the same behavior. Thus power profiles/patterns within a cluster are represented by the centroid, which is the mean of the patterns within the cluster, and the number of patterns in the cluster, which represents the density of the cluster. Based on the observation that outliers are minority, we assume that a cluster of high number of patterns represents a normal power consumption behavior. On the other hand, a cluster of low number of patterns, is likely an outlier/abnormal power consumption behavior. This enables us to define and localize outliers in the SG, which enhances the reliability of the SG. Then, the representative data (*i.e.*, centroid and number of patterns) of the clusters is communicated with the central processor of the SG, in order to find global power profiles. This would reduce the communication cost because only the representative data is sent to the central processor. In addition, the overall efficiency of the system will be improved as the central processor analyzes the representative data only.

The second level of data clustering takes place at the central processor where the centroids of the data clusters derived at the first level constitute a new dataset. Then global power profiles are derived by clustering the new dataset (*i.e.*, dataset of the centroids of the local power profiles). The global power profiles are represented by the centroids of the clusters at the second clustering level weighted by the number of power patterns in each local power profile.

1) Local Power Profiles Generation (Level 1): The K-Means is a well-known and widely used clustering algorithm because of its simplicity and ease of implementation. In this work, we use the K-means algorithm in both levels of data clustering of the proposed model, because of its minimal computational cost when compared to other models. In addition, the complexity of the proposed model can easily be defined with respect to the parameters of model, and thus different setups can easily be compared.

Let $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N$ be a set of datapoints in d-dimensional space, where d is the number of features, and K is a predefined number of clusters. The K-means algorithm minimizes the objective function given by :

$$F(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N) = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in c_k} \|\mathbf{x}_i - \bar{\mathbf{x}_k}\|^2, \tag{1}$$

where c_k denotes the k^{th} cluster,

$$\bar{\mathbf{x}_k} = \frac{1}{n_k} \sum_{\mathbf{x}_i \in c_k} \mathbf{x}_i, \tag{2}$$

is the center of the k^{th} cluster, and n_k is the number of datapoints in k^{th} cluster. $\|.\|$ denotes the Euclidean norm used by the K-means algorithm. The algorithm starts with K datapoints that represent the centroids of the clusters. Each datapoint in the dataset is assigned to the centroid of the closest cluster and the mean of the datapoints in the same cluster is calculated. The procedure is repeated iteratively until convergence or the exit condition is satisfied.

Let $X_l = \{\mathbf{x}_{(l,1)}, ..., \mathbf{x}_{(l,N_l)}\}$ be the available dataset of layer l, which includes the row data vectors of N_l residences. First, X_l is clustered into K_l mutually exclusive clusters. Let $C_l = \{\mathbf{c}_{(l,1)}, ..., \mathbf{c}_{(l,K_l)}\}$ be the set of the centroids of the resultant clusters in layer l, where $\mathbf{c}_{(l,j)}$ denotes the centroid of the j^{th} cluster in the l^{th} layer and K_l is the number of clusters in

layer l. Note that by using K-means, the number of clusters K_l must be predetermined. Here, we use the same value of K_l for all layers (*i.e.*, K). Also, $n_l = \{n_{(l,1)}, ..., n_{(l,K_l)}\}$ denotes the set of the number of users in each cluster, where $n_{(l,j)}$ is the number of users in the j^{th} cluster of the l^{th} layer. Next, C_l and n_l are transmitted to the centralized processor of the SG in order to derive the global data profiles.

2) Local Power Profiles Generation (Level 2): Let $X' = \{\mathbf{x}'_{(1)}, ..., \mathbf{x}'_{(L\times K)}\}$ be the available dataset at the central processor. X' contains the centroids of the clusters of all layers derived in the first clustering level where $\mathbf{x}'_{(1)} = \mathbf{c}_{(1,1)}$, $\mathbf{x}'_{(K)} = \mathbf{c}_{(1,K)}$, and $\mathbf{x}'_{(L\times K)} = \mathbf{c}_{(L,K)}$ where L is the number of layers. Then the K-means algorithm is used again to partition the data X' into K' clusters. After the convergence of this clustering process, the derived centroids in C' do not properly describe the global power profiles, since each local centroid, which is derived at the first clustering level, represents different number of residences (*i.e.*, $n_{(l,i)}$). To this end, the global energy consumption patterns are defined as:

$$\mathbf{c}'_{(j)} = \frac{\sum_{l,i:\mathbf{c}_{(l,i)}\in cluster j} \mathbf{c}_{(l,i)} \times n_{(l,i)}}{\sum_{l,i:\mathbf{c}_{(l,i)}\in cluster j} n_{(l,i)}}, \forall j = 1, ..., K'. \quad (3)$$

Note that the proposed multi-layered clustering model is general and can easily be modified to adopt different clustering algorithms in each level. Fig. 2 illustrates the proposed multi-layered clustering model.

Implementation of the proposed model, by using the K-means clustering algorithm, and discussion of its complexity are described in the following subsection.

3) Complexity Analysis: Usually, distributed clustering is used to reduce the communication demand. It is possible to reduce the computational complexity of the system by using the proposed multi-layered approach, which increases profitability.

In general, the complexity of the K-means algorithm is given by:

$$Complexity = \mathcal{O}(N \times K \times P \times I), \quad (4)$$

where N is the number of $P-$dimensional datapoints/vectors in the dataset, K is the number of clusters, and I is the number of iterations until convergence. Let X_l be the dataset of the l^{th} layer and I_l number of iterations at that layer, the complexity of the K-means clustering at each layer of the first level is given by:

$$Complexity_{(l)} = \mathcal{O}(N_l \times K \times P \times I_l). \quad (5)$$

Since the first clustering level includes multiple layers, the overall complexity of the first clustering level is given by:

$$Complexity_{(1^{st}Level)} = \mathcal{O}(\sum_{l=1}^{L} N_l \times K \times P \times I_l). \quad (6)$$

In the special case that all layers have the same number of datapoints (*i.e.*, residences), the overall complexity of the first clustering level can be written as:

$$Complexity_{(1^{st}Level)} = \mathcal{O}(L \times N_l \times K \times P \times I_{max}), \quad (7)$$

where I_{max} denotes the maximum allowable number of iterations.

Because the clustering process at each layer is independent of other layers, they can be executed in parallel. Thus, the worst case execution time of the first level can be given by:

$$E - Time_{(1^{st}Level)} = \mathcal{O}(N_l \times K \times P \times I_{max}). \quad (8)$$

The worst case computational complexity of the second clustering level is given by:

$$Complexity_{(2^{nd}Level)} = \mathcal{O}((L \times K) \times K' \times P \times I_{max}), \quad (9)$$

which coincides with the execution time of the second level.

Therefore, the overall worst case complexity of both levels, when all layers have the same N_l and K_l, is given by

$$Complexity_{(Total)} = \mathcal{O}(((N_l \times K) + (L \times K) \times K') \times (P \times I_{\max})). \quad (10)$$

On the other hand, in the case that the centralized K-means is used to find the clusters of users power profiles, the corresponding computational complexity is given by:

$$Complexity_{(Centralized)} = \mathcal{O}(N \times K' \times P \times I_{max}), \quad (11)$$

which is higher than $Complexity_{(Total)}$ when

$$L < \frac{N \times K' - N_l \times K}{K \times K'}. \quad (12)$$

Thus, the approach discussed in this section can also be used for reducing the computational complexity if the condition stated in (12) is satisfied.

IV. Experiments and Results

In this section, we describe the dataset used to evaluate the performance of the proposed model and present the results of the experiments.

A. Dataset

Building a consumer power profiling model requires a dataset from which the model learns. The dataset shall also represent real data that describes real-world scenario. In our experiments, the proposed model is evaluated on the well-referenced UMass Smart Microgrid Data Set [10], [26].

The UMass Smart Microgrid Data Set was gathered by the Smart project. This dataset contains average electricity usage data from 400 anonymous homes in western Massachusettes, USA at one minute granularity for an entire day. For privacy reasons, the data source and the homes are kept anonymous. This data is well-suited for emulating microgrids or examining the grid-scale effects of various optimizations, such as the use of energy storage [27].

Excluding the inactive buildings, we focused on 395 residential homes. For the needs of the simulation setup, the dataset is separated into L datasets, where the l^{th} dataset is processed by the l^{th} layer at the first level of the proposed model.

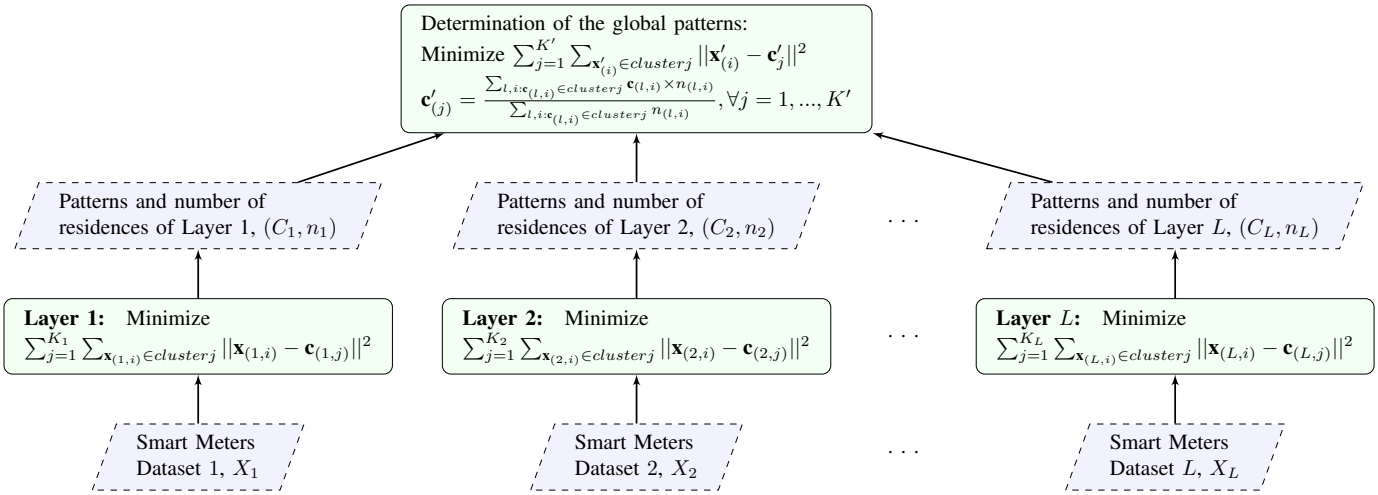


Fig. 2. Proposed multi-layered clustering model.

TABLE I
SIMULATION SETUP

Variable	Description	Value
Number of local patterns in each layer	K	3
Maximum number of iterations	I_{max}	100
Number of layers	L	8
Number of global patterns	K'	5

B. Performance Evaluation and Results Analysis

To show the effectiveness of the proposed model, in this subsection, we provide a detailed simulation of the proposed model on the UMass Smart Microgrid Data Set. The parameters of the proposed model are given in Table I. Following [10], the number of global patterns (*i.e.*, K'), has been set to 5, while K has been selected according to (12). The selection of the value of L is a trade-off between the effectiveness of the clustering process and its computational complexity. This is shown in the next subsection.

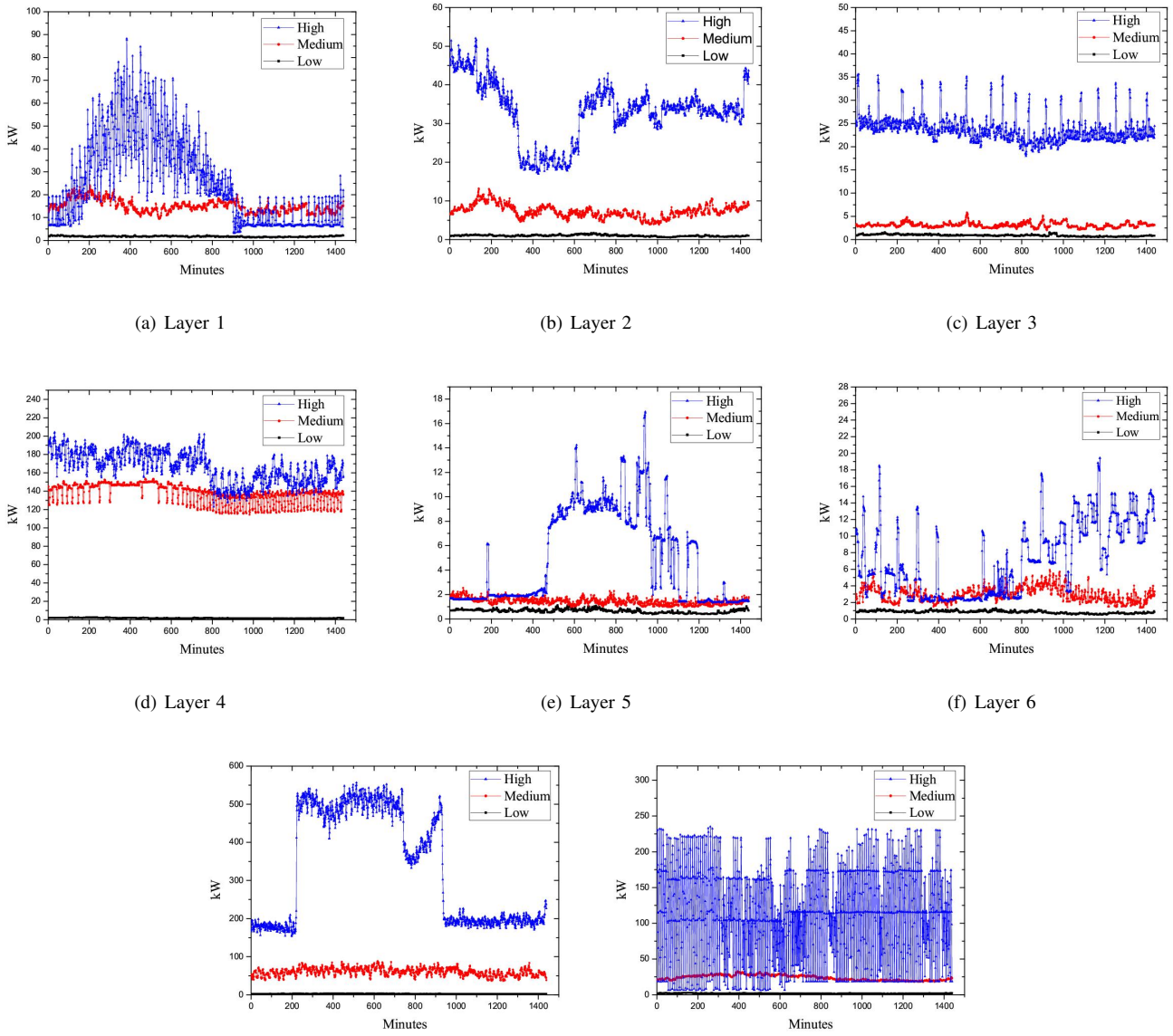
In order to have a better understanding of the significance of the layered approach, we studied the local data patterns at the different layers of the proposed model. Fig. 3 depicts the derived local patterns (*i.e.*, the centroids at each layer). In each layer, the locally available power consumption dataset is partitioned into three clusters: i) low consumption, ii) medium consumption, and iii) high consumption cluster. As can be seen in Fig. 3, the derived local data patterns at different layers are distinct. For example, a power consumption that is characterized as a high-consumption in layer 2 corresponds to a low or medium consumption in layer 4. Of particular interest, the peak period differs from a layer to another. For example, the behavior of the per-residence power consumption characterized as high in layer 1 reaches its peak in the period of (200–800) minutes. On the other hand, the behavior of the per-residence power consumption characterized as high in layer 2 reaches its minimum during the same period. This means that the knowledge of local data patterns is important as it allows us to develop an efficient and effective location-aware pricing scheme as well as better scheduling algorithms.

As mentioned earlier, the derived local data patterns are partitioned into K' homogeneous clusters, where each cluster represents a global data pattern. The global data patterns are characterized as: i) low, ii) low-medium , iii) medium, iv) high, and v) very high consumption. Then the derived global data patterns are compared to the derived patterns when the centralized K-means approach is used (*i.e.*, the whole power consumption data set is collected and clustered in a single location). As can be observed in Fig. 4, the derived patterns using both methods are similar. This motivates the use of the proposed model as a promising alternative approach of the centralized K-means clustering as it maintains the quality and reduces the computational complexity of power profile clustering. More specifically, the reduction of the overall complexity, when this setup (number of instances=395, $L = 8$, $K = 3$, and $K' = 5$) is used, is:

$$R = 1 - \frac{(395 \times 3 \times +8 \times 3 \times 5\times) \times P \times I_{max}}{395 \times 5 \times P \times I_{max}} = 33.92\% \tag{13}$$

More importantly, as it can be observed from Fig. 3 and Fig. 4, the global patterns characterized as high consumption and very high consumption are provoked by the power profiles initially processed by layers 8 and 7, respectively. Thus, the proposed model allows some type of power consumption outliers localization, while preserving individual residences privacy.

1) The Trade-Off between Complexity and Performance: To show the effectiveness of the proposed multi-layered power consumption profiling model, we compared its performance with the performance of the fully centralized K-means. We ran 10^3 independent simulations for each setup (*i.e.*, L value). This is because the K-means is a local minimizer algorithm that tries to find the global optima; however, it is not necessarily find the global optimum. This is because the resultant clusters of K-means algorithm depend heavily on the initialization parameters (*i.e.*, seed, centroids). Different simulation with different initialization parameters might lead to different clustering results.



(a) Layer 1 (b) Layer 2 (c) Layer 3

(d) Layer 4 (e) Layer 5 (f) Layer 6

(g) Layer 7 (h) Layer 8

Fig. 3. Local patterns.

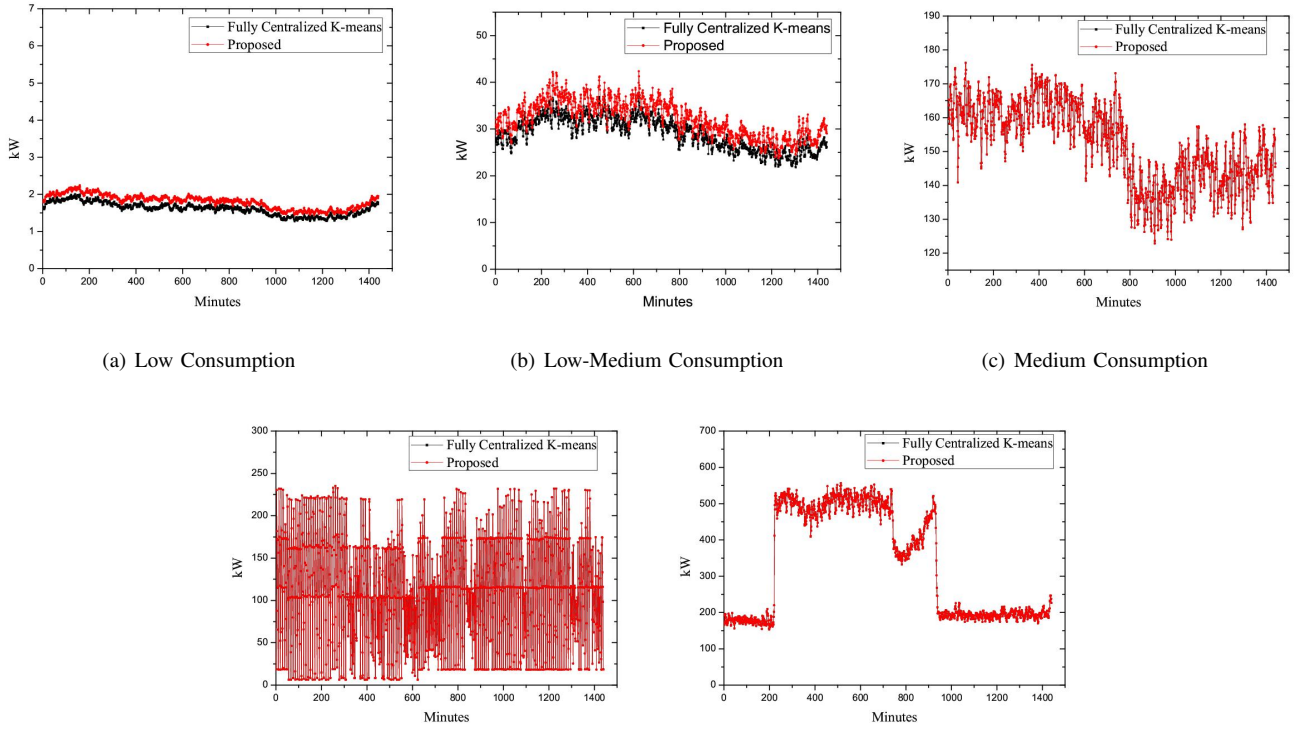
Several evaluation measures were used in the literature to quantify the performance of clustering algorithms. Most of these measures consider how well the clusters are separated. However, a good clustering algorithm should consider the density of the clusters. Thus we have selected silhouette coefficient [28] to compare the performance of the proposed model and the centralized K-means. Silhouette coefficient is defined as:

$$S = \frac{\sum S(i)}{N}, \tag{14}$$

where

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{15}$$

where i is a datapoint in the dataset, $a(i)$ is the average distance of datapoint i to the other datapoints in the same cluster as i, and $b(i)$ is the minimum average distance of datapoint i to datapoints in other clusters. As $a(i)$ measures how dissimilar i is to its own cluster, the smaller $a(i)$ value is, the more compact the cluster is. The value of $b(i)$ implies the degree of difference between i and other clusters, thus the larger $b(i)$ is, the more separated i is from other clusters. The value of the silhouette coefficient is between -1 and 1. A positive silhouette coefficient value means the cluster including i is compact and i is far from other clusters, while negative silhouette coefficient value means i is closer to the datapoints in another cluster than to the datapoints in its own cluster. Normally, for silhouette coefficient value, the bigger, the better, and value 1 is the extreme preferable condition [28]. Note that when the proposed model is used, we consider that i follows its local centroid to global clustering. Thus, two datapoints



(a) Low Consumption (b) Low-Medium Consumption (c) Medium Consumption

(d) High Consumption (e) Very High Consumption

Fig. 4. Global patterns.

represented by the same local centroid, are placed in the same global cluster.

Fig. 5 shows the performance of the proposed multi-layered clustering model, while varying the number of layers (*i.e.*, L), compared to the fully centralized K-means clustering. As can be seen in Fig. 5, the performance of the proposed multi-layered model increases as the number of layers increases. The reason for this is that for a small value of L (*e.g.*, $L = 4$), where the value of K and the number of datapoints are fixed, the number of datapoints in each layer is high; thus the clusters centroids might not adequately capture the spacial properties of the data. In addition, as the number of the representative centroids is a function of the number of layers and K, for a small value of L, the number of representative centroids is also small. Therefore, in general, increasing the number of layers would increase the quality of data clustering.

In addition to the quality of clustering, complexity is another important issue to consider especially when dealing with large-scale data (*e.g.*, smart grid data). Fig. 6 shows the reduction in complexity of the proposed model when varying the number of layers. In this context, we have studied the reduction of complexity of three cases: i) the worst-case overall computational complexity, ii) the real overall computational complexity, and iii) the communication complexity, which corresponds to the power profiles transmission between the local aggregators and the central processing unit of the SG. As can be seen in Fig. 6, the proposed model considerably reduces the computational and communication complexity for the whole range of the number of layers(*i.e.*, $4 \leq L \leq 20$). Both of the worst-

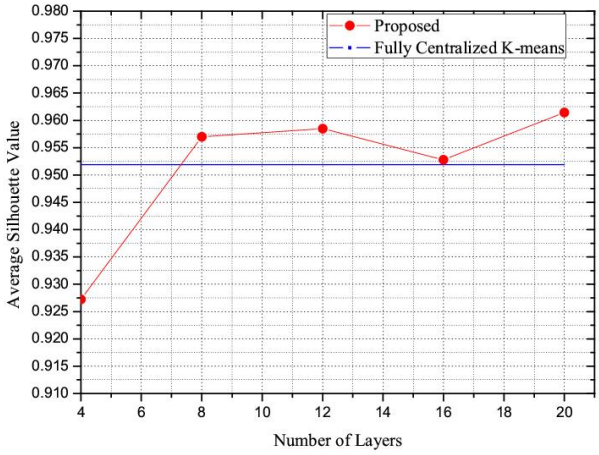


Fig. 5. Average silhouette value comparison.

case computational and communication complexity of the proposed model also increases with the number of layers. This is because as the value of L increases the number of the local patterns that have to be sent to the central processing unit of the SG also increases. However, it is remarkable that the complexity reduction of the overall computational complexity of the proposed model is higher than the one of the worst-case computational complexity. This is because, as the number of the power profiles clustered by each layer reduces, the

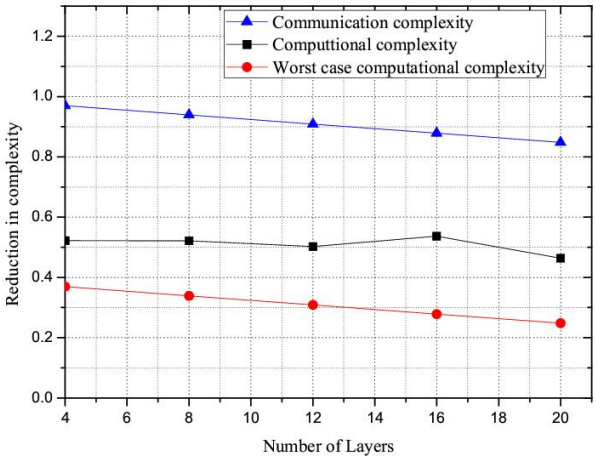


Fig. 6. Reduction in computational and communication complexity.

number of iterations (*i.e.*, I_l) required for the local K-means to converge reduces also. Similarly, as the number of layers (*i.e.*, L) increases, the number of the power profiles clustered by each layer reduces, which requires less number of iterations for the local K-means to converge. Thus, the real increase of the complexity of the proposed model is not necessarily linear with respect to the number of layers.

Figs. 6 and 5 jointly show that the performance and the complexity of the proposed model are highly related. For example, when $L = 20$ both performance and complexity of the proposed model present a local maxima. Interestingly, the proposed model performs at least as well as the centralized K-means for $L \geq 8$, for which value the reduction in computational complexity is about 52% and the reduction in the communication complexity is about 95%. Finally, for specific values of L, the proposed model seems to perform slightly better than the centralized one. This might be because of the independent processing of the power profiles among different layers, which allows avoiding local optimal solutions.

V. Conclusion

This paper presents a multi-layered clustering model for power consumption profiling of consumers in SGs. The proposed model aims to reduce the communication and computational complexity of the power consumption profiling process, which is essential for constructing an effective prediction, pricing, and anomaly detection models in end-users level. The proposed model consists of two levels of clustering. In the first level, the concept of layers is introduced where a layer is defined as an instance of K-means clustering that operates on data aggregated from a certain region. The data of a layer is portioned into mutually exclusive clusters. Customers power consumption patterns in a cluster are represented by the centroid (*i.e.*, local power consumption patterns) and the number of patterns within the cluster. The second level partitions the data (*i.e.*, local power consumption patterns of all layers) generated in the first level into multiple clusters. The centroids of the clusters of the second level weighted by the number of patterns in each cluster in the first level represent the global power consumption patterns. Experiments results show that the proposed model can significantly reduce the communication and computation cost of the power consumption profiling process while maintaining the performance, outperforming the centralized power consumption profiling approach. This makes the proposed model a preferable candidate for power consumption profiling in SGs especially when dealing with large-scale data.

Our future work includes investigating how to combine the advantages of the proposed method with advanced features selection/extraction methods [10], studying on how the proposed model could be used to construct more accurate power consumption prediction models considering customer power consumption clustering both at the local and the global levels [3]. For an example, the dataset concerns 400 residential houses for one day. The current study has no verified analytic results on consumption behaviours based on day, month or season. Furthermore, considering the proposed two-levels clustering offers some type of outliers localization, it could be used in the context of local marginal prices, where the total power consumption can be smoothed by only altering the local electricity prices. In terms of the clustering, it is also worth to compare the partition quality with others clustering algorithms. Finally, this work could be extended to meet the requirements of real-time data processing applications, such as clustering the power consumption of appliances and possibly detecting of appliances with anomaly behavior such as faulty or compromised appliances.

Acknowledgment

This publication was made possible with the support of ICT Fund. The statements made herein are solely the responsibility of the authors.

References

[1] M. Erol-Kantarci and H. T. Mouftah, "Energy-efficient information and communication infrastructures in the smart grid: A survey on interactions and open issues," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 1, pp. 179–197, 2015.

[2] E. Span, L. Niccolini, S. D. Pascoli, and G. Iannacconeluca, "Last-meter smart grid embedded in an internet-of-things platform," *IEEE Transactions on Smart Grid*, vol. 6, no. 1, pp. 468–476, Jan. 2015.

[3] P. Goncalves Da Silva, D. Ilic, and S. Karnouskos, "The impact of smart grid prosumer grouping on forecasting accuracy and its benefits for local electricity market trading," *IEEE Trans. Smart Grid*, vol. 5, no. 1, pp. 402–410, Jan 2014.

[4] S.-l. Yang, C. Shen *et al.*, "A review of electric load classification in smart grid environment," *Renewable and Sustainable Energy Reviews*, vol. 24, pp. 103–110, 2013.

[5] Y. Wang, Q. Chen, C. Kang, M. Zhang, K. Wang, and Y. Zhao, "Load profiling and its application to demand response: A review," *Tsinghua Science and Technology*, vol. 20, no. 2, pp. 117–129, 2015.

[6] A. Grandjean, J. Adnot, and G. Binet, "A review and an analysis of the residential electric load curve models," *Renewable and Sustainable Energy Reviews*, vol. 16, no. 9, pp. 6539 – 6565, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1364032112004820

[7] K. le Zhou, S. lin Yang, and C. Shen, "A review of electric load classification in smart grid environment," *Renewable and Sustainable Energy Reviews*, vol. 24, pp. 103 – 110, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1364032113001822

[8] A. Shahzadeh, A. Khosravi, and S. Nahavandi, "Improving load forecast accuracy by clustering consumers using smart meter data," in *2015 International Joint Conference on Neural Networks (IJCNN)*, Jul. 2015, pp. 1–7.

[9] M. N. Macedo, J. J. Galo, L. A. Almeida, and A. C. Lima, "Typification of load curves for {DSM} in brazil for a smart grid environment," *International Journal of Electrical Power & Energy Systems*, vol. 67, pp. 216 – 221, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0142061514007200

[10] A. K. Marnerides, P. Smith, A. Schaeffer-Filho, and A. Mauthe, "Power Consumption Profiling using Energy Time-Frequency Distributions in Smart Grids," *IEEE Commun. Lett.*, vol. 19, no. 1, pp. 46–49, Jan. 2015.

[11] H. Yang, L. Zhang, Q. He, and Q. Niu, "Study of power load classification based on adaptive fuzzy c means," *Power System Protection and Control*, vol. 38, no. 16, pp. 112–115, 2010.

[12] P. R. Jota, V. R. Silva, and F. G. Jota, "Building load management using cluster and statistical analyses," *International Journal of Electrical Power & Energy Systems*, vol. 33, no. 8, pp. 1498 – 1505, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0142061511001438

[13] N. K. Visalakshi and K. Thangavel, *Foundations of Computational, IntelligenceVolume 6: Data Mining*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, ch. Distributed Data Clustering: A Comparative Analysis, pp. 371–397.

[14] S. F. Bush, S. Goel, and G. Simard, "IEEE vision for smart grid communications: 2030 and beyond roadmap," *IEEE Std. Association*, pp. 1–19, Sept 2013.

[15] Y. Yamagata, H. Seya, and S. Kuroda, "Smart Community Clustering for Sharing Local Green Energy," in *Proc. International Conference and Utility Exhibition on Green Energy for Sustainable Development (ICUE)*, Mar. 2014, pp. 1–7.

[16] A.-H. Mohsenian-Rad, V. W. S. Wong, J. Jatskevich, R. Schober, and A. Leon-Garcia, "Autonomous demand-side management based on game-theoretic energy consumption scheduling for the future smart grid," *IEEE Trans. Smart Grid*, vol. 1, no. 3, pp. 320–331, Dec 2010.

[17] P. Samadi, H. Mohsenian-Rad, V. W. S. Wong, and R. Schober, "Tackling the Load Uncertainty Challenges for Energy Consumption Scheduling in Smart Grid," *IEEE Trans. Smart Grid*, vol. 4, no. 2, pp. 1007–1016, Jun. 2013.

[18] R. Li, Q. Wu, and S. S. Oren, "Distribution Locational Marginal Pricing for Optimal Electric Vehicle Charging Management," *IEEE Trans. Power Systems*, vol. 29, no. 1, pp. 203–211, Jan. 2014.

[19] H. Narimani and H. Mohsenian-Rad, "Autonomous Demand Response in Heterogeneous Smart Grid Topologies," in *Proc. IEEE PES Innovative Smart Grid Technologies (ISGT)*, Feb. 2013, pp. 1–6.

[20] X. Fang, S. Misra, G. Xue, and D. Yang, "Smart gridthe new and improved power grid: A survey," *IEEE communications surveys & tutorials*, vol. 14, no. 4, pp. 944–980, 2012.

[21] G. Chicco, R. Napoli, F. Piglione, P. Postolache, M. Scutariu, and C. Toader, "Comparisons among clustering techniques for electricity customer classification," *IEEE TRANSACTIONS ON POWER SYSTEMS PWRS*, vol. 21, no. 2, p. 933, 2006.

[22] A. Mutanen, M. Ruska, S. Repo, and P. Jarventausta, "Customer classification and load profiling method for distribution systems," *IEEE Transactions on Power Delivery*, vol. 26, no. 3, pp. 1755–1763, 2011.

[23] M. Piao, H. S. Shon, J. Y. Lee, and K. H. Ryu, "Subspace projection method based clustering analysis in load profiling," *IEEE Transactions on Power Systems*, vol. 29, no. 6, pp. 2628–2635, 2014.

[24] G. J. Tsekouras, N. D. Hatziargyriou, and E. N. Dialynas, "Two-stage pattern recognition of load curves for classification of electricity customers," *IEEE Transactions on Power Systems*, vol. 22, no. 3, pp. 1120–1128, 2007.

[25] P. P. Rodrigues and J. Gama, "Holistic distributed stream clustering for smart grids," in *Proc. Workshop on Ubiquitous Data Mining*, 2012, pp. 18–22.

[26] "Smart dataset," http://traces.cs.umass.edu/index.php/Smart/Smart, Accessed: 2016.04.05.

[27] S. Barker, A. Mishra, D. Irwin, E. Cecchet, P. Shenoy, and J. Albrecht, "Smart*: An open data set and tools for enabling research in sustainable homes," *SustKDD, August*, vol. 111, p. 112, 2012.

[28] Z. Wang and R. S. Srinivasan, "Classification of Household Appliance Operation Cycles: A Case-Study Approach," *Energies*, vol. 8, no. 9, pp. 10 522–10 536, 2015.

Omar Y. Al-Jarrah received the B.S. degree in Computer Engineering from Yarmouk University, Jordan, in 2005, the M.S. degree in Engineering from The University of Sydney, Sydney, Australia in 2008 and the Ph.D. degree in Electrical and Computer Engineering from Khalifa University of Science and Technology, United Arab Emirates, in 2016. Currently he is working as a Postdoctoral Fellow in the Department of Electrical and Computer Engineering, Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates. His main research interest involves machine learning, intrusion detection, big data analytics, and knowledge discovery in various applications.

Yousof Al-Hammadi is currently a Director of Graduate Studies and Assistant Professor at the Electrical & Computer Engineering Department, Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates. He received his Bachelor degree in Computer Engineering from KUSTAR (previously known as Etisalat College of Engineering), UAE in 2000, MSc degree in Telecommunications Engineering from the University of Melbourne, Australia in 2003, and PhD degree in Computer Science and Information Technology from the University of Nottingham, UK in 2009. His main research interests are in the area of Information security which includes Intrusion Detection, Botnet/Bots Detection, Viruses/Worms Detection, Artificial Immune Systems, Machine learning, RFID Security and Mobile Security.

Paul D. Yoo (M'11-SM'13) is currently with the Cranfield Defence and Security, based in Defence Academy of the United Kingdom, Shrivenham. Prior to this, Dr. Yoo held academic/research posts in Sydney, Bournemouth, and the UAE. Dr. Yoo serves as Editor of IEEE COMML and Elsevier JBDR journals and holds over 60 prestigious journal and conference publications. Dr. Yoo is affiliated with University of Sydney and Korea Advanced Institute of Science and Technology (KAIST) as Visiting Professor. He is a Senior Member of IEEE and a Member of BCS. His research focuses on large-scale data analytics including design and development of computational models and algorithms inspired by intelligence found in physical, chemical and biological systems, and to solve practical problems in security and digital forensics.

Sami Muhaidat Muhaidat received the Ph.D. degree in Electrical and Computer Engineering from the University of Waterloo, Ontario, in 2006. From 2007 to 2008, he was an NSERC postdoctoral fellow in the Department of Electrical and Computer Engineering, University of Toronto, Canada. From 2008 to 2012, he was an Assistant Professor in the School of Engineering Science, Simon Fraser University, BC, Canada. He is currently an Assistant Professor at Khalifa University of Science and Technology and a Visiting Reader in the Faculty of Engineering, University of Surrey, UK. Samis research focuses on advanced digital signal processing techniques for image processing and communications, machine learning, cooperative communications, vehicular communications, MIMO, and space time coding. He has authored more than 100 journal and conference papers on these topics. Sami is an active Senior IEEE member and currently serves as an Editor for IEEE Communications Letters and an Associate Editor for IEEE Transactions on Vehicular Technology. He was the recipient of several scholarships during his undergraduate and graduate studies. He was also a winner of the 2006 NSERC Postdoctoral Fellowship competition..