World Academy of Science, Engineering and Technology
International Journal of Energy and Power Engineering
Vol:13, No:3, 2019

Consumer Load Profile Determination with Entropy-Based K-Means Algorithm

Ioannis P. Panapakidis, Marios N. Moschakis

***Abstract*—**With the continuous increment of smart meter installations across the globe, the need for processing of the load data is evident. Clustering-based load profiling is built upon the utilization of unsupervised machine learning tools for the purpose of formulating the typical load curves or load profiles. The most commonly used algorithm in the load profiling literature is the K-means. While the algorithm has been successfully tested in a variety of applications, its drawback is the strong dependence in the initialization phase. This paper proposes a novel modified form of the K-means that addresses the aforementioned problem. Simulation results indicate the superiority of the proposed algorithm compared to the K-means.

***Keywords*—**Clustering, load profiling, load modeling, machine learning, energy efficiency and quality.

I. Introduction

LEAST-COST electric system planning includes methods to deal with the increasing demand. Contrary to the traditional system planning where the demand needs are covered via a proportional expansion of the generation capacity, in least-cost planning the aim is to utilize tools to modify the demand patterns in order to postpone or eliminate the requirement to install new generation units. Demand Side Management (DSM) is a basic tool of least-cost planning; it refers to a family of measures that target at the reduction or shifting of the demand from peak to off-peak periods. DSM is seen as a way to manage the demand so that network congestion, capacity shortage and investments to expensive power generation technologies would be avoided [1]. DSM is built upon two pillars, namely energy efficiency and demand response. Especially with demand response programs, the consumer becomes more active in competitive energy markets. Therefore, the benefits of DSM, apart from utilities and system operators, are visible to the end-consumers also [2]-[4].

For the purpose to fully materialize the benefits of DSM, accurate knowledge of the demand patterns is required. Load data measurements are important in order to design, implement, and evaluate a DSM program. Smart metering installations are increasing across the globe [5]. Intelligent metering infrastructure provides the benefit of time-interval data collection. After the data collection, the processing phase takes place for the purpose of information retrieval and knowledge extraction.

Clustering is an unsupervised machine learning tool with proven robustness in a variety of information retrieval and pattern recognition problems. A clustering algorithm is completely data-driven; no prior information is necessary regarding the optimal number of clusters. Thus, a clustering algorithm should be executed for variable number of clusters. When the clustering error is minimized, the optimal number of cluster is determined [6].

During the last years, a large number of researches have tested clustering algorithms in various load datasets. The scope is to group together load curves with similar shapes and extract the representative load curves or load profiles of the load sets under study. Clustering-based load profiling involves the implementation of clustering algorithms for the purpose of formulating the load profiles of sole consumers and consumer clusters [7]. The load profiles can be utilized in load forecasting tasks, DSM applications and others.

The importance of the clustering-based load profiling technical field is reflected by the large variety of algorithms that have been proposed and tested. The algorithms can be classified in the following categories: i) Partitional algorithms such as the K-means, K-medoids and others, ii) Hierarchical algorithms, such as the average linkage, the single linkage and others, iii) Fuzzy algorithms, such as Fuzzy C-Means, iv) neural network-based algorithms, such as the Self-Organizing Map, the Hopfield Network and v) algorithms that do not belong to the previous categories, such as Support Vector Clustering, Modified "Follow-the-Leader" and others. In most studies, a comparison takes place between algorithms of different type. This is common approach followed in the literature; there is no universally acclaimed clustering algorithm. Different algorithms lead to the best results in different clustering problems and set-ups [8].

K-means is the most popular algorithm in the load profiling literature. This is due to the algorithm's speed, minimal complexity, comprehensive operation and software availability. The operation of the algorithm is centered in a cost minimization iterative process. The main drawback of the algorithm is its strong dependence of the initial conditions, i.e. the initial centroids or cluster centers are selected randomly. Therefore, a poor selection leads accordingly to poor clustering results.

The present paper proposes a modified version of the K-means, namely, the "Entropy K-means". The scope is to reach out into an optimal selection of the initial centroids that upgrade the operation of the algorithm. A comparison with the conventional K-means takes place and the superiority of the

I. P. Panapakidis is with the University of Thessaly, 41110, Larissa, Greece (e-mail: ipanap@ee.auth.gr)

M. N. Moschakis is with the University of Thessaly, 41110, Larissa, Greece (phone: +302410684325; fax: +302410684325; e-mail: mmoschakis@teilar.gr).

proposed clustering algorithm is reported.

II. Load Profiling Framework

A. Pattern Representation

The term "pattern" refers to the input of the clustering algorithm. The pattern representation stage refers to the selection of the technique that will express the patterns for further processing. In the present study, the pattern corresponds to the daily active load curve. We deal with *n*=3 consumers of different type, namely a low voltage residential, a low voltage commercial and a medium voltage industrial. For each consumer $m = 1,...,M$ with *M*=365 daily load curves are available. The pattern of the *m*-th consumer is denoted as $p^{(m)} = \{P_1^{(m)},...,P_D^{(m)}\}$ where P refers to the mean active load value and *D* is the dimension of the pattern. The dimension *D* equals to 24 or 96, if the load measurements are taken in 1 h or 15 min intervals, respectively. Moreover, we denote the minimum and the maximum value of $p^{(m)}$ as $x_{\min}^{(m)} = \min\{p^{(m)}\}$ and $x_{\max}^{(m)} = \max\{p^{(m)}\}$, respectively. Since the load profiling process deals with the similarity of the patterns` shapes and not with the patterns` magnitudes, all values are normalized in the [0,1] range according to the following expression:

$$x^{(m)} = \{x_1^{(m)},...,x_D^{(m)}\} = \frac{p_i^{(m)} - x_{\min}^{(m)}}{x_{\max}^{(m)} - x_{\min}^{(m)}} \quad (1)$$

where $i = 1,2,...,D$. The set of the patterns is denoted as $X = \{x^{(m)}, m = 1,...,M\}$. The clustering process is a mapping of $M \to K$, where K is the number of clusters and $1 \le K \le M$.

Each formulated cluster has a centroid which is the average pattern of all patterns that belong to the cluster. The centroid is also expressed by a*D*-dimensional vector:

$$c^{(k)} = \{c_1^{(k)},...,c_D^{(k)}\} = \frac{1}{M_K} \sum_{\substack{m=1 \\ x^{(m)} \in C_k}}^{M} x^{(m)} \quad (2)$$

where M_K is the number of vectors that belong to the cluster C_k. The set of the clusters is denoted as $C_k = \{c^{(k)}, k = 1,...,K\}$. The load profiles are the weighted sum of the load pattern data that belong to a cluster C_k,

$$lp_k = \{lp_{k1},...,lp_{kD}\} = \sum_{\substack{m=1 \\ x^{(m)} \in C_k}}^{M} [x^{(m)}(x_{\max}^{(m)} - x_{\min}^{(m)}) + x_{\min}^{(m)}] \quad (3)$$

The output of the clustering procedure is the extraction of the centroids and the cluster labels, i.e. the membership of patterns in the clusters.

B. Clustering Algorithms

The *K*-means algorithm has been successfully applied in a variety of applications. It is a favorable method in many areas of application due to its simplicity and linear complexity, which is defined as $O(I*n*K*D)$, where I is the number of iterations, n is the number of input features, K is the number of clusters and D is the dimension of the features. Normally, $(I,K) \ll n$. K-means is the most commonly used partitional clustering algorithm. The scope of partitional clustering is to simply divide a set of patterns into non-overlapping subsets or clusters, such that each feature belongs to exactly one subset [10]. The main concern is the determination of K initial centroids, one for each cluster. The outcome of the procedure is highly dependent on the location of these centroids in the feature space. The K-means algorithm consists of the following steps:

Step1. Initialization. Start the algorithm with a random selection of K centroids from the subset $C_k \subset X$.

Step2. Clustering. At each iteration *t* define each C_{tk} as: For each $j = 1,...,i$ assign $x_j \in C_{ik}$ where k is chosen so that $\left\|x_j - c_t^{(k)}\right\| = \min_{1 \le l \le K} \left\|x_j - c_t^{(l)}\right\|$, $j = 1,...,t$.

Step3. Updates. The new centroids of each cluster are calculated as $c_t^{(k)} = \frac{1}{M_k^t} \sum_{x_j \in C_{tk}} x_j$, $\forall k \in \{1,...,k\}$ where M_k^t is the population of set C_{tk} during iteration *t*.

Step4. Termination. The algorithm stops if there is no change in the partition at the *t*-th iteration; otherwise we increment *t*to *t*+1and repeat Steps 2 and 3.

The selection of the initial centroids is critical to the algorithms performance. Since the performance is heavily influenced by the choice of the initial centroids, the main disadvantage of the K-means is that the several executions of the algorithm lead to different results, i.e. different populations of the clusters and hence, different final centroids. This is the reason for the algorithm` s converge to local optima, i.e. sub-optimal solutions. While the number of clusters is not a priori determined, the algorithm should be executed with variant number so that to reach the optimal solution.

In the present paper, the aforementioned problem is addressed by selecting patterns with specific complexity. The Shannon Entropy is used to quantify the complexity of the load time series of each consumer. In the paper, the complexity of the time series is approached as the level of volatility. The Shannon Entropy *H* is given by [9]:

$$H = -\sum_i p_i \log p_i \quad (4)$$

where p_i is the probability of the value *i* to appear in the time series. Following this concept, a pattern (i.e. daily load curve) with high *H* contains hourly load values that are repeated, i.e. they are present more than one instant with the daily time period. Therefore, the volatility is relatively low. On the other hand, entropy values close to 0, refer to low probabilities of repeated values and thus, time series with high degree of non-linearity. Thus, in order to create initial clusters that are distant

in the feature space, patterns with large variations in entropy values should be selected. Actually, the entropy value provides an indication about the shape of the load curve. The more distant are the initial clusters, the better initial partitioning of the patterns is accomplished. Fig. 1 provides an example of 3 daily load curves with different entropy values. The load curves are normalized in the [0,1] range correspond to the small commercial consumer. The load curve of Fig. 1 (a)) presents many fluctuations compared to the one of Fig. 1 (c)). They correspond to different entropy values and therefore, they are quite dissimilar in terms of shapes. This means that they can serve as two different initial centroids in the initialization step of the K-means.

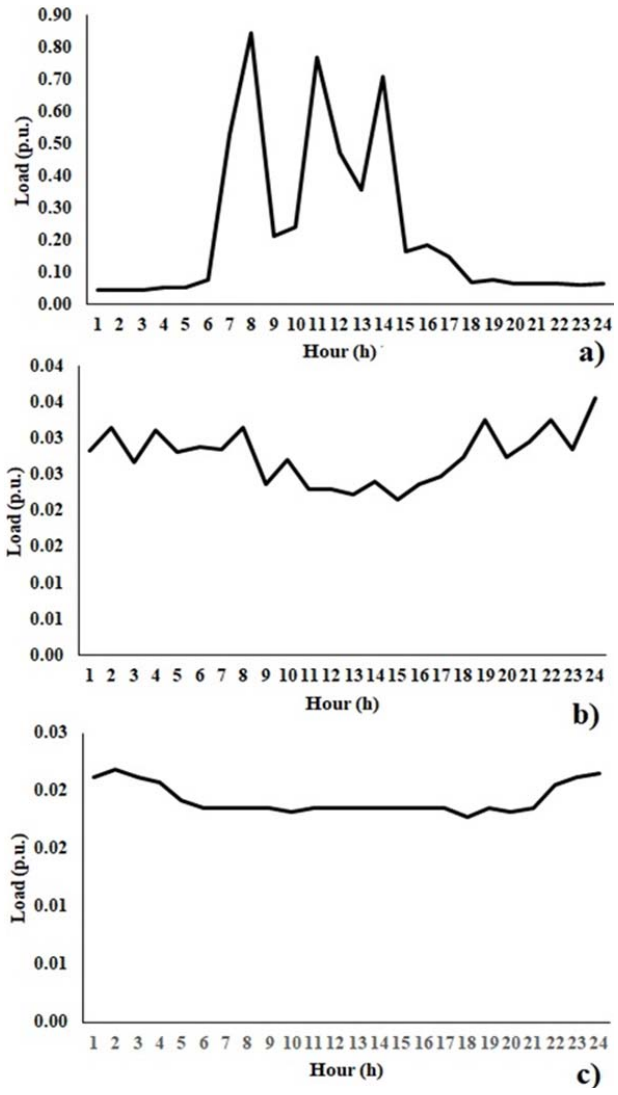

Fig. 1 Example of three patterns with entropy values equal to: a) 2.29, b) 4.08 and c) 4.58

The flow-chart of the proposed Entropy K-means algorithm is illustrated in Fig. 2. For each consumer, the Shannon Entropy is calculated per pattern. Next, the list with the entropy values is sorted. If the desirable number of clusters is k=2, the patterns that refer to the first and last entropy values of the list are selected. The quantiles of the sorted list are regarded for k>2. Excluding the initialization step, the proposed algorithm is similar to the K-means.

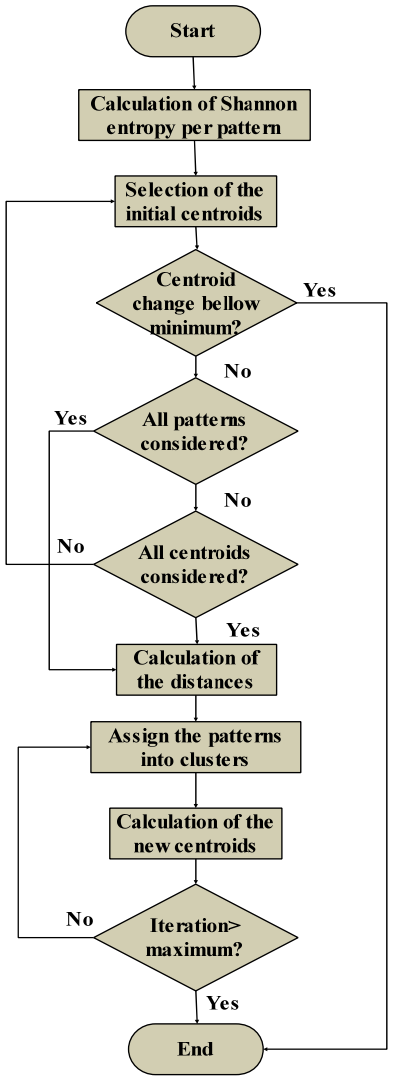

Fig. 2 Flow-chart of the operation of the proposed algorithm

C. Clustering Validation

Clustering validation involves the utilization of a set of indicators that assess the algorithms` performance. The validation indicators are built upon similarity metrics. The score of the metrics indicates the similarity of among the patterns. Prior to the presentation of the indicators, the following metrics are defined [11]:

i) The Euclidean distance between two vectors $x^{(s)}$ and $x^{(t)}$, with $(x^{(s)}, x^{(s)}) \in X$, is:

$$d(x^{(s)}, x^{(t)}) = \sqrt{\frac{1}{D}\sum_{h=1}^{D}\left(x_h^{(s)} - x_h^{(t)}\right)^2} \quad (5)$$

ii) The subset of X that belong to the C_k cluster is denoted as S_k. The Euclidean distance between the centroid $c^{(k)}$ of

the *k*-th cluster and the subset S_k is the geometric mean of the Euclidean distances $d\left(c^{(k)}, S_k\right)$ between $c^{(k)}$ and each member $x^{(k)}$ of S_k :

$$d\left(c^{(k)}, S_k\right) = \sqrt{\frac{\sum_{i=1}^{M_k} x^{(k)} \in S_k d^2\left(c^{(k)}, x^{(k)}\right)}{M_k}} \quad (6)$$

iii) The geometric mean of the inner-distances between the features members of the subset S_k is

$$d\left(S_k\right) = \sqrt{\frac{1}{2M_k} \sum x^{(k)} \in S_k d^2\left(x^{(k)}, x^{(m)}\right)} \quad (7)$$

The Clustering Dispersion Indicator (CDI), which is the ratio of the mean infra-set distance between the input vectors in the same cluster and the infra-set distance between the clusters` centroids,

$$\mathrm{CDI} = \frac{\sqrt{\frac{1}{K}\sum_{k=1}^{K} d^2\left(S_k\right)}}{\sqrt{\frac{1}{2K}\sum_{K=1}^{K} d^2\left(c_k, C_k\right)}} \quad (8)$$

The Inter cluster Index (IEI) is the dispersion between the clusters and is defined as:

$$\mathrm{IEI} = \sum_{k=1}^{K} M_k \cdot (c^{(k)} - p) \cdot (c^{(k)} - p)^t \quad (9)$$

where $p = \frac{1}{M}\sum_{m=1}^{M} p^{(m)}$ is the arithmetic mean of the input vectors.

The Intra cluster Index (IAI) is the dispersion within the same cluster and is defined as:

$$\mathrm{IAI} = \sum_{m,k=1}^{K} \sum_{x^{(m)} \in X} (x^{(k)} - c^{(k)}) \cdot (x^{(k)} - c^{(k)})^t \quad (10)$$

In the present study, three validity indicators are taken into account in order to provide a robust assessment framework. The assessment should regard a set of indicators that measure different cluster qualities.

III. Simulation Results

The comparisons are shown in Figs. 3-5. The scope of the validation framework is through a comparative analysis to examine the robustness of the Entropy K-means over the conventional version of the algorithm. We considered a large variation range for the number of clusters in order to examine the potential of the Entropy K-means for many clusterings. No prior information about the number of clusters is available. Also, no data pre-processing stage took place, such as de-trending, outlier removal and others. The algorithms are executed separately for each consumer for 2 to 30 clusters and for each number the values of the indicators are checked. Thus, we are referring to three comparisons, one for every consumer. In order to keep the comparison fair and accurate, the number of iterations for both algorithms is set to t =500.

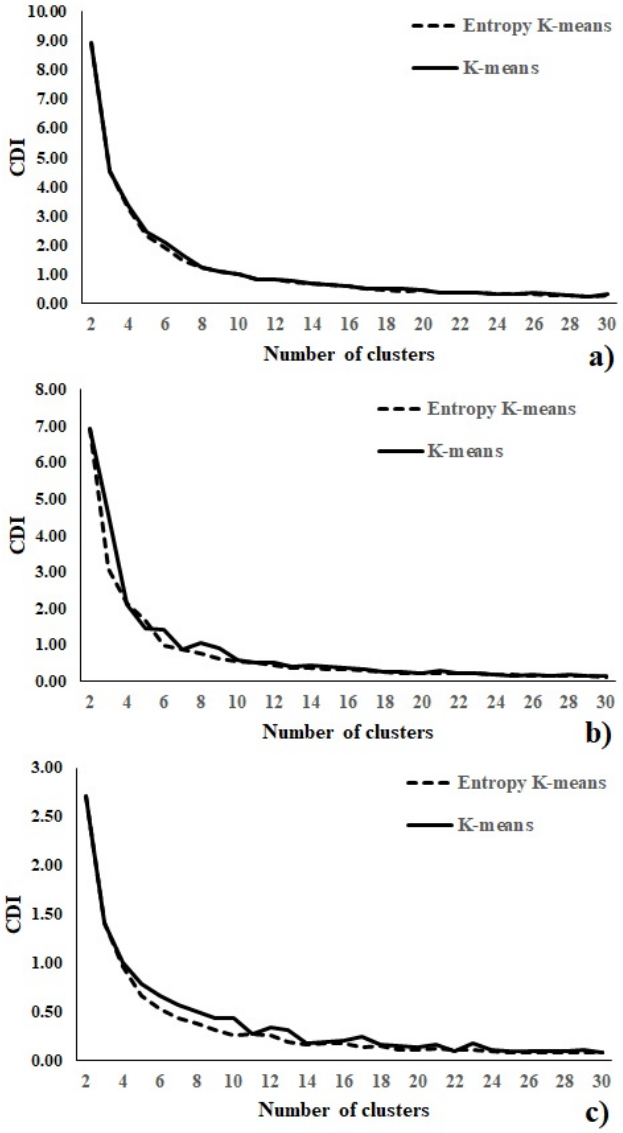

Fig. 3 Comparison of the algorithms using the CDI indicator considering the: a) 1st, b) 2nd and c) 3rd consumer, respectively

The CDI is a measure of the compactness and the separation of the clusters. The compactness is a degree that shows how close are the patterns of the same clusters themselves and how close they are to the centroid of the cluster they belong. The separation refers to the distances between the centroids of the different clusters.

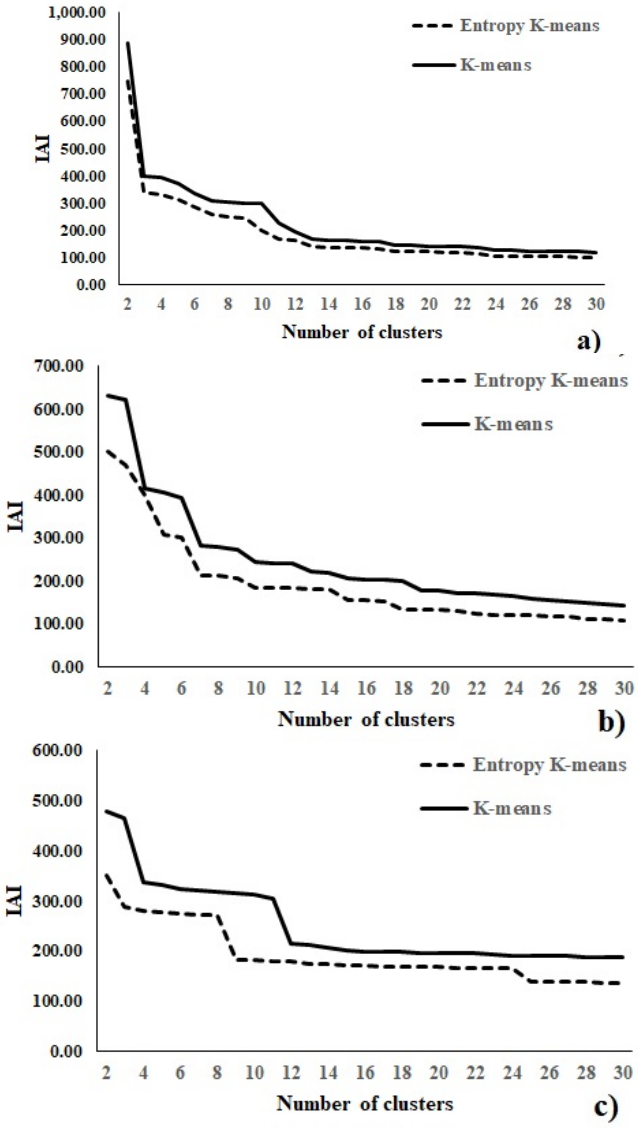

Fig. 4 Comparison of the algorithms using the IAI indicator considering the: a) 1st, b) 2nd and c) 3rd consumer, respectively

A robust clustering method should lead to low values of compactness and large values of separation, i.e. the patterns should be close to each other in the feature space and the clusters, as represented by their centroids, should be well separated. While the number of clusters is increasing, the CDI receives lower values.

According to Fig. 3, the proposed Entropy K-means leads to better clusterings in all cases. It should be noted that the superiority of an algorithm over the other should refer to low values in most of clusters. This is a fact not only in the case of the CDI but to the IAI also. Using this indicator, the difference between the two algorithms is more visible. The IAI is another measure of compactness. Contrary to the CDI, the shape of IAI do not present strict monotonically decreasing tendency but provides a more visible indicator regarding the algorithms comparison.

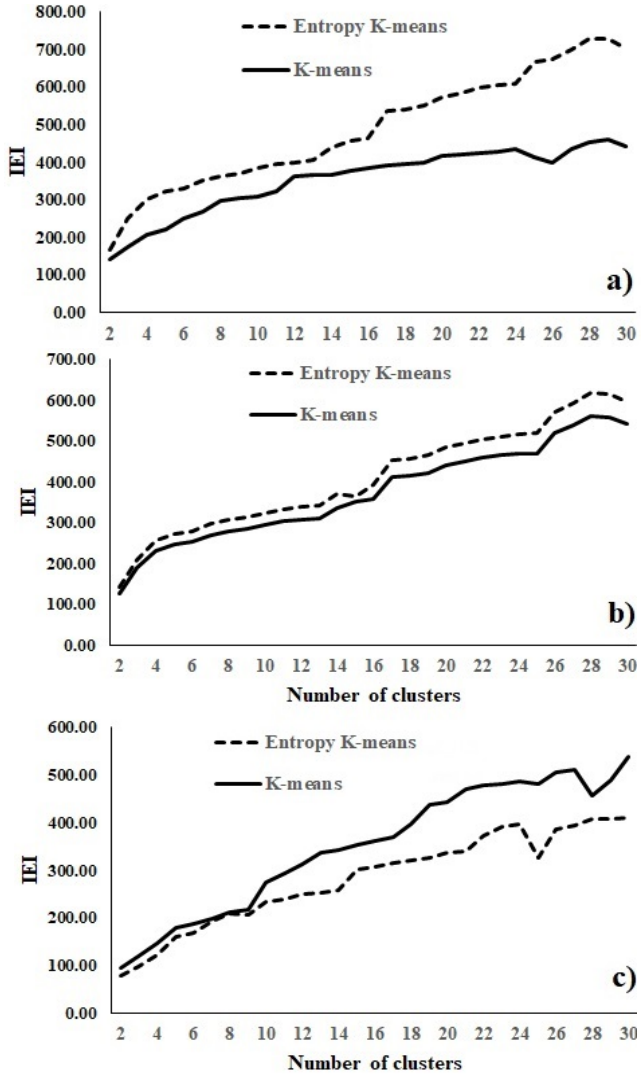

Fig. 5 Comparison of the algorithms using the IEI indicator considering the: a) 1st, b) 2nd and c) 3rd consumer, respectively

The IEI receives larger values while the number of clusters is increasing. Here, the superior algorithm should lead to higher values. While the number of cluster is increasing, the patterns are becoming more distant from the overall mean. Again, the proposed algorithm outperforms the K-means in all cases.

While the shape of the IEI indicator displays many fluctuations, it is not recommended to be used for determining the optimal number of clusters. This is not the case for the CDI and IAI. For extracting the optimal number of clusters, the CDI has been selected since it is a measure for two cluster qualities. By applying the "knee" point detection method on the CDI curve produced by the Entropy K-means, the optimal number of clusters for each consumer is drawn [12]. More specifically, the optimal numbers for the 1st, 2nd and 3rd consumers are 12, 12 and 11, respectively. The load profiles of the consumers are shown in Fig. 6. The 1st consumer has profiles that present considerable difference between them. There are some profiles that correspond to low consumption.

The corresponding clusters are mainly composed by holidays and weekends. This is also the case for the 2nd consumer. The 3rd consumer is a medium voltage industry. The load profiles are less elastic to the factors that influence them.

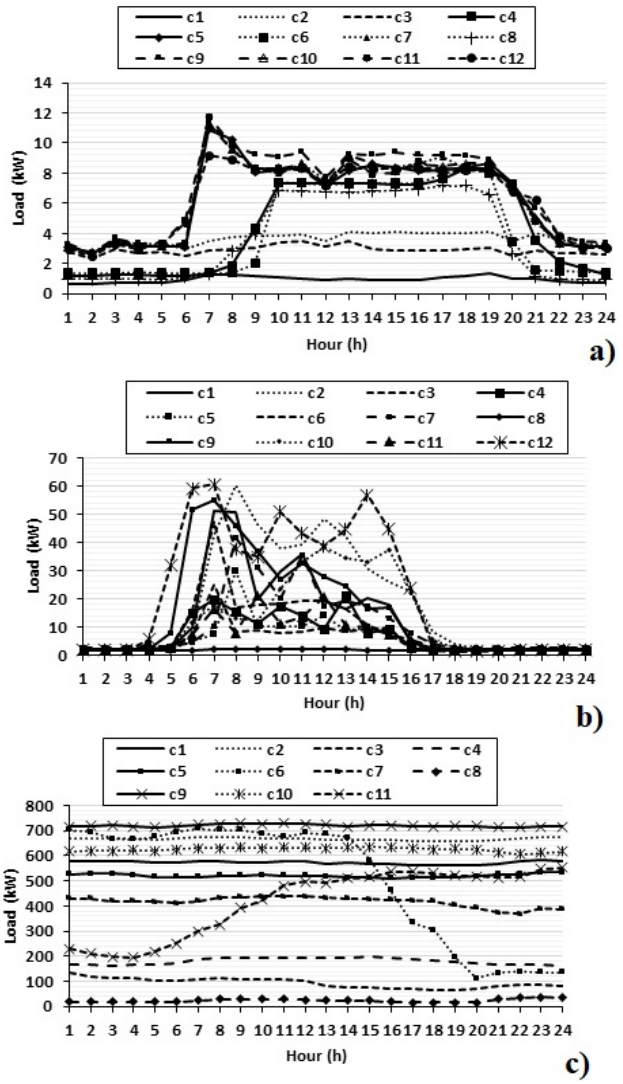

Fig. 6 Load profiles of the: a) 1st, b) 2nd and c) 3rd consumer, respectively

IV. Conclusion

Clustering-based load profiling is an important technical field within the power systems community. The benefits of formulating the load profiles of various consumers are evident in the implementation of DSM measures and other applications. This fact is recognized by the researchers and resulted in the investigation of many algorithms to address the issue of clustering load curves and extracting the load profiles. Furthermore, clustering is viewed as a way to process big amounts of data, a concept known as "Big Data".

The present paper contributes to the load profiling related literature by presenting a novel clustering algorithm that leads to lower errors compared to the K-means, the most common algorithm in the literature. The Entropy K-means deals with the drawback of the random selection of patterns that will serve as initial centroids. The Shannon Entropy is calculated per pattern. The selection of the centroids is held considering the patterns that display high variances of entropy values, i.e. they are highly dissimilar one another. The algorithms have been compared with a set of validity indicators. All indicators denote the superiority of the proposed algorithm.

The present analysis will be further expanded by comparing the Entropy K-means with other algorithms in the literature and considering more validity indicators. Also, other load data sets will be used to fully examine the potential of the proposed algorithm.

References

[1] Z. Hu, X. Han, and Q. Wen, "Integrated Resource Strategic Planning and Power Demand-Side Management", 1st ed., Springer-Verlang: Berlin, Germany, 2013, pp. 63-133.

[2] F. Rahimi and A. Ipakchi, "Overview of demand response under the smart grid and market paradigms", In Proc. of the 2010 IEEE Innovative Smart Grid Technologies Conference, 19-21 January, Gaithersburg, Maryland, USA, pp. 1-7.

[3] International Energy Agency (IEA), Technology Roadmap Smart Grids, IEA: Paris, France, 2011.

[4] H.T. Haider, O.H. See, W. Elmenreich, "A review of residential demand response of smart grid", Ren. Sustain. Energy Rev. vol. 59, 2016, pp. 166-178.

[5] S.S.S.R. Depuru, L. Wang and V. Devabhaktuni, "Smart meters for power grid: Challenges, issues, advantages and status", Ren. Sust. Energy Rev., vol. 15, no. 6, pp. 2376-2742, August 2011.

[6] R. Xu and D. Wunsch, Clustering, Hoboken New Jersey, John Wiley & Sons Inc., 2006.

[7] G. Chicco, R. Napoli and F. Piglione, "Comparisons among Clustering techniques for electricity customer classification", IEEE Trans. Power Syst. vol. 21, no.2, May 2006, pp. 933-940.

[8] G. Chicco, "Overview and performance assessment of the clustering methods for electrical load pattern", Energy, vol. 42, 2012, pp. 68-80.

[9] X. Liu, A. Jiang, N. Xu and J. Xue, "Increment entropy as a measure of complexity for time series", Entropy, vol. 18, no. 1, pp. 1-12.

[10] D. Steinley,"K-means clustering: A half-century synthesis", British Journal of Mathematical and Statistical Psychology, vol. 59, 2006, pp. 1-34.

[11] 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, August 2007, pp. 1120-1128.

[12] Q. Zhao, M. Xu and P. Fränti, "Knee point detection on Bayesian information criterion", In Proc. of the 20th IEEE International Conference on Tools with Artificial Intelligence 2008, 3-5 November 2008, Dayton, Ohio, USA, pp. 431- 438.