A data-mining-based methodology to support MV electricity customers' characterization

Sérgio Ramos, João M. Duarte, F. Jorge Duarte, Zita Vale

A B S T R A C T

This paper presents an electricity medium voltage (MV) customer characterization framework supported by knowledge discovery in database (KDD). The main idea is to identify typical load profiles (TLP) of MV consumers and to develop a rule set for the automatic classification of new consumers. To achieve our goal a methodology is proposed consisting of several steps: data pre-processing; application of several clustering algorithms to segment the daily load profiles; selection of the best partition, corresponding to the best consumers' segmentation, based on the assessments of several clustering validity indices; and finally, a classification model is built based on the resulting clusters. To validate the proposed framework, a case study which includes a real database of MV consumers is performed.

Keywords:
Load profiling
Data mining
Clustering
Classification
Clustering validity

1. Introduction

Concerning the electricity market environment, the characterization of the electrical consumers assumes an important supporting tool for electric utilities, for understanding and predicting the behaviour of their electricity customers. It is expected for suppliers to know, as much as possible, the electrical consumption habits of their customers, to offer them suitable electric energy services at the least cost and thus differentiating themselves from others competitors. The knowledge about customers' consumption patterns is particularly important for setting up dedicated commercial offers. Indeed, load patterns are broadly used in tariff design, system planning, system maintenance, load management and marketing [1]. Typically, the electrical suppliers companies cluster consumers into representative classes and use the representative load profile to study consumers' behaviour [2–5].

Automatic meter reading (AMR) systems, normally operating at quarter-hour intervals, have been implemented by most of electric power companies [6], mainly for MV customers. In fact, the European Union's current strategy promotes its utilization [7]. So, gradually, a huge amount of data concerning electricity consumption will become available and stored into databases, allowing load patterns to be extracted from these. In the deregulated electricity industry there is a distinct separation throughout the value chain of the power system: production, transmission, distribution and retail. While transmission and distribution companies explore the distribution power network, according different voltage levels, the retail companies are responsible for managing relations with end consumers, including invoicing, billing and customer services, and have some flexibility in formulating the tariff offers, assuring that their offers meet the requirements by the regulatory authorities in the form of prices [5,8].

Conceptually, the tariffs offers are formulated with reference to a specific consumer's class, defined by a set of technical and commercial attributes. The distinction between customers' groups can be made based on the definition of macro-categories, e.g., residential, commercial, industries, public lighting, or others specific consumers. There are also other attributes that can be used for the distinction of customers' classes, such as the contracted power value, annual energy consumption and the voltage level or, as presented in [9], a criterion based on the cost of energy purchased from the pool market by a retailer. However, in large number of research works in this field of study, the load profile for tariffs purpose is typically performed with load data. Also, the customers' characterization could be accomplished, for example, based on the commercial type of activity. However, the load profiles that belong to the same commercial type of activity reveal different electrical consumption habits. Thus, using the commercial type of activity for customers' categorization is generally not efficient for representing

the electricity consumption [4,5,8]. For the electricity customers without measured data available, their association with one of the formed typical load profile classes can be identified *à posteriori* based on available information and attributes of that customer and of the obtained customer classes. Power utilities can also obtain load profiles from AMR customers and the so-called virtual load profile (VLP) from non-AMR customers in order to create load profile of all customers [5,6].

In the last years, dedicated research effort has been developed in order to study load profiling. Typically, pattern recognition methods have been applied to electricity consumption data. A variety of clustering algorithms have been proposed to group together load diagrams with similar shapes. In [10] it is possible to find a brief overview of well-known clustering methods, discussing its major challenges and some of the emerging and useful research directions are pointed out.

This paper presents a data-mining-based methodology to identify typical load profiles, using a real database provided by the Portuguese utility. To conduct data partitioning, several clustering algorithms have been used and the evaluation of the quality of the obtained data partitions were assessed by cluster validity indices. The implemented methodology is extremely useful for electrical suppliers' companies, as well as consumers' aggregators, to identify the typical daily load profile supporting the design of new tariff structures and to improve their strategy of market share, either by optimizing their power purchase option, or by the definition of demand response programs. A new customer can easily be placed in one of the defined clusters using a classification model. With a significant increase of clients, one may need to start the whole clustering process to find the new optimal data partition, which can be seen as a limitation of the proposed approach.

The remaining of this paper is organized as follows. In Section 2 a review of data mining techniques is presented. Section 3 addresses the proposed methodology for electrical customers' characterization and classification. In Section 4 a case study using real data is presented. The last section summarizes the concluding remarks.

2. Data mining techniques

Data mining is the task of discovering patterns in large data sets involving methods of artificial intelligence, machine learning, statistics, and database systems. In this section, a brief description of some methods used for data clustering analysis and classification is presented.

2.1. Data clustering algorithms

Clustering is the process of partitioning a set of data objects into clusters based on a concept of similarity or proximity among data. Even though there is a huge number of clustering algorithms in the literature [11,12], no single algorithm can effectively find by itself all types of cluster shapes and structures.

The purpose of any clustering technique consists in dividing a data set *Xcomposed* of n data patterns $\{x_1, \ldots, x_n\}$ into K clusters $\{C_1, \ldots, C_K\}$, such that similar data patterns are placed in the same cluster and dissimilar data patterns are grouped into different clusters. The set of clusters $P = \{C_1, \ldots, C_K\}$ is referred as data partition. The major clustering algorithms can be classified into the following categories:

I. Partitive algorithms initially define K seed points $\bar{x}_k$ (centroids or medoids), one for each cluster, and iteratively update these points to optimize some objective function. At each iteration, each object x_i is assigned to the most similar seed point. Three **partitive** algorithms are shortly described ahead.

The K-Means algorithm [13] is the best known data clustering algorithm. K-Means tries to minimize the within-cluster sum of squares $\left(\left(\sum_{k=1}^{k}\sum_{x_i \in C_k}\left\|x_i - \bar{x}_k\right\|^2\right)\right)$ where $\left\|x_i - \bar{x}_k\right\|^2$ is the Euclidean distance between pattern x_i and its closest cluster centroid $\bar{x}_k$.

This algorithm takes as parameter the desired number of clusters K and randomly chooses K data patterns as the initial centroids $\{\bar{x}_1, \ldots, \bar{x}_K\}$ of each cluster. Then, K-Means algorithm iterates between two steps: find for each pattern $x_i \in X$ the closest centroid $\bar{x}_k$ and assign it to the corresponding cluster C_k, and update each centroid $\bar{x}_k$ as the mean vector of the corresponding cluster. This process is repeated until no pattern assignments are changed from one iteration to the next one, meaning the algorithm converged to a (local) minimum.

Clustering process can be made given some constraints between data patterns (pairwise constrained clustering). The PC-KMeans algorithm (PCKM) [14] formulates the goal of clustering in the pairwise constrained clustering framework as minimizing a combined objective function, which is defined as the sum of the total square distances between the points and their cluster centroids (like in K-Means) and the cost of violating the pairwise constraints (must-link and cannot-link constraints between data patterns).

The MPC-KMeans algorithm (MPCKM) [15] is an extension of the PC-KMeans algorithm by proposing the incorporation of a metric learning directly into the clustering algorithm in a way that allows pairwise constraints to influence the metric learning process along with pairwise constraints. Basically the MPC-KMeans algorithm combines the objective function of the PC-KMeans algorithm with the learning of the distance metric.

II. Hierarchical algorithms create a hierarchical decomposition of a given set of data objects. A hierarchical algorithm can be agglomerative or divisive, based on how the hierarchical decomposition is formed. The agglomerative approach starts with each object forming a separate cluster and in each successive iteration merges the objects or clusters that are close to one another, until all of the clusters are merged into one, or until a termination condition holds. The divisive approach starts with all of the objects in the same cluster and in each successive iteration a cluster is split up into smaller clusters, until eventually each object is in one cluster, or until a termination condition holds. The single-link, average-link and complete-link [16] are examples of agglomerative algorithms.

III. Density-based algorithms [17] consider high-density regions in space as clusters, and objects in low-density regions as outliers or noise. Their general idea is to continue growing clusters as long as the density (number of objects or data points) in the "neighbourhood" exceeds some threshold.

IV. Grid-based algorithms [18] quantize the object space into a finite number of cells obtained by splitting each data feature into intervals. These cells form a grid structure. Clusters are formed by finding contiguous cells containing a minimum number of objects.

V. Spectral algorithms use the highest K eigenvalues to build a new representation of data. Then, a fast clustering algorithm, such as K-Means, is applied to perform clustering on the new representation.

The Normalized Cut algorithm [19] transforms the clustering algorithm into a weighted graph partitioning problem $G = (V, E)$, such that the vertices of the graph $V = \{v_1, \ldots, v_n\}$ correspond to the data patterns and the weights w_{ij} for each edge $E = \{e_{ij} : 1 < i < n-1, 2 < j < n, i < j\}$ correspond to the similarity between a pair of data patterns, and partitions the graph into K clusters.

VI. Model-based algorithms [20] assume a model for each cluster and find the best fit of the data to the given model. They locate clusters by constructing a density function that reflects the spatial distribution of the data points.

2.2. Clustering validity indices

While clustering data, there are two important questions that must be addressed: how many clusters are present in the data and how good is the data partition itself. Clustering validity indices provide the formal mechanisms to give an answer to these questions. There are a wide number of clustering validity measures [21] but none get the best result in all data sets. Here we present some of the most representative.

Assuming a_i as the average distance between $x_i \in C_l$ and the other objects in the same cluster, and b_i as the minimum average distance between x_i and all objects grouped in another cluster, the silhouette width is defined for each object x_i as $s_i = (b_i - a_i)/\max\{a_i, b_i\}$, and indicates how well x_i is adjusted to its cluster when compared to other clusters. The **Silhouette index** (S) [22], is given by the average silhouette width computed over all objects in the data set, $S = 1/n\sum_{i=1}^{n} s_i$.

The **Hubert's Statistic** (H) [23] measures the correlation between a $n \times n$ co-membership matrix, U, representing the data partition P, and a $n \times n$ distance matrix D, with the distances between all pairs of objects. The co-membership $U = [U_{ij}]$ is built by setting each entry U_{ij} to 0 if both x_i and x_j was assigned to the same cluster or to 1 otherwise. The Hubert's Statistic is defined as

$$H(P) = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} U_{ij} D_{ij},$$

considering the matrices to be symmetric. A high $H(P)$ value indicate a good data partition, however $H(P)$ values increases with the number of clusters. A **Normalized version of Hubert's Statistic** (NH) prevents this bias and is defined as

$$NH(P) = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{(U_{ij} - \mu_U)(D_{ij} - \mu_D)}{\sigma_U \sigma_D}$$

where μ_U, μ_D, μ_U and μ_D are the means and standard deviations of U and D, respectively.

Dunn index (D) [24] is defined as

$$\mathrm{Dunn}(P) = \min_{i=1,\ldots,k} \left\{ \min_{j=i+1,\ldots,k} \left(\frac{d_{c_i,c_j}}{\max_{l=1,\ldots,k} \mathrm{diam}(C_l)} \right) \right\}$$

where d_{c_i,c_j} is a dissimilarity function between two clusters C_i and C_j defined as $d_{c_i,c_j} = \min_{x_a \in C_i, x_b \in C_j} d_{x_a,x_b}$. The diameter of a cluster is defined as $diam_{C_l} = \max_{x_a, x_b \in C_1} d_{x_a,x_b}$ and can be considered as a measure of clusters' dispersion. A high value of this index indicates compact and well separated clusters.

Consider s_{C_i,C_j} a similarity measure between two groups C_i and C_j, based on a measure of dispersion of a group C_i $\left(\mathrm{disp}_{c_i}\right)$ and a measure of dissimilarity between two groups C_i and C_j $(d_{c_i c_j})$, $s_{c_i c_j} = \left(disp_{c_i} + disp_{c_j}\right)/d_{c_i c_j}$. The index **Davies–Bouldin index** (DB) [25] is defined as

$$DB(P) = \frac{1}{K} \sum_{i=1}^{K} s_{C_i}$$

where $s_{C_i} = \max_{i,j=1,\ldots,K, i \neq j} s_{C_i C_j}$. As DB corresponds to the average similarity between each group of the data set to its most similar group, low values of DB indicate good partitions.

SD index [21] is based on the concepts of average dispersion of groups and total separation between groups. The average dispersion of groups is defined as

$$Disp = \frac{1}{K} \sum_{i=1}^{K} \frac{\left\|\sigma(C_i)\right\|}{\left\|\sigma(X)\right\|}$$

where $\sigma(C_i)$ is the variance in the $i^{\underline{o}}$ group and $\sigma(X)$ is the variance of the entire data set. Full separation of the groups is given by

$$Sep(k) = \frac{D_{\max}}{D_{\min}} \sum_{q=1}^{K} \left(\sum_{l=1}^{K} \left\|C_q - C_l\right\| \right)^{-1}$$

where $D_{\max} = \max(||C_i - C_j||)\,\forall\, i,\ j \in \{1, 2, \ldots, K\}$ is the greatest distance between the centers of two groups, and $D_{\min} = \min(||C_i - C_j||)\,\forall\, i,\ j \in \{1, 2, \ldots, K\}$ is the shortest distance between the centers of two groups. The SD index is defined as $SD = a \times Disp \times Sep(K)$ where $a = Sep(K_{\max})$ and $K_{\max}$ is the maximum value to be considered for the number of groups K.

Consider a matrix M with $K \times n$ elements representing a data partition of the data set into K groups where $m_{kj} = 1$ if the object $x_j \in C_k$, otherwise $m_{kj} = 0$. The I **index** [26] is defined as

$$I = \left(\frac{1}{K} \times \frac{E_1}{E_K} \times D_K \right)^p$$

where $E_K = \sum_{i=1}^{K} \sum_{j=1}^{n} m_{ij} ||x_j - c_i||$ and $D_K = \max_{1 \le i,j \le K} ^{||C_i - C_j||}$. The correct value for the number of groups is the value of K for which I is maximized. Generally p takes the value 2.

Point Symmetry Index (PS) [27] measures the symmetry average of the data objects relatively to the centers of the groups to which they belong. The center of each group (c_i) is obtained by

$$c_i = \frac{\sum_{j=1}^{n} \left(S_{ij}\right)^m x_j}{\sum_{j=1}^{n} \left(S_{ij}\right)^m}$$

where S_{ij} are entries of a matrix S that represents a data partition and m is a user specified parameter. In the case of a crisp data partition, S_{ij} corresponds to 1 or 0 depending on whether or not the object x_j belongs to the group C_i. m is only used in the case of a fuzzy data partition. The Point Symmetry index is defined as

$$PS = \frac{1}{K} \sum_{i=1}^{K} \left[\frac{1}{n_i} \sum_{j=1}^{n_i} \frac{dc_{x_j,c_i}}{d_{\min}} \right]$$

where $d_{\min}$ is the minimum Euclidean distance between two centers of the groups and $dc_{x_j,c_i} = ds_{x_j,c_i} . d_{x_j,c_i}$. The distance ds_{x_j,c_i} is calculated by

$$ds_{x_j,c_i} = \min_{1 \le l \le n_i, l \neq j} \left\{ \frac{\left\|(x_j - c_i) + (x_l - c_i)\right\|}{\left\|(x_j - c_i)\right\| + \left\|(x_l - c_i)\right\|} \right\}$$

and dc_{x_j,c_i} is the Euclidean distance between x_j and c_i. The lowest value of PS indicates the best data partition of the data set.

The **XB index** [28] is defined as

$$XB = \frac{\sum_{i=1}^{K} \sum_{j=1}^{n} S_{ij}^2 \left\|c_i - x_j\right\|^2}{n \times \min_{i,l} \left\|c_i - c_l\right\|^2}$$

where s_{ij} is defined as in the Point Symmetry index. Low values of this index indicate compact and well-separated groups.

2.3. Classification algorithms

Unlike the data clustering techniques, the classification algorithms are supervised, as the class of each object in the data set is known *a priori*. The objective of these algorithms consists on learning a function or a set of rules, denominated as classifier, which allows assigning the correct class for a new (unobserved) object. There are several types of algorithms to train classifiers which can be organized by learning strategy:

- **Statistical models** assume the classes of objects are generated according to some probabilistic distribution. Examples are the linear and quadratic discriminant analysis [29];
- **Classification tree** algorithms learn a tree structure such that all non-leaf nodes are attributed a feature, and each leaf node is labelled with a class. Starting in the root node, at each node, the data is successively split by identifying which feature, and splitting threshold, better discriminates classes in the corresponding subset of data [30]. For instance, the C5.0 algorithm uses the information gain criterion;
- **Artificial neural networks** (ANN) emerged with the intention of mathematically modeling the human brain. The most common type of ANNs is the Multilayer perceptron, generally trained using the back propagation algorithm [31];
- **Support vector machines** algorithms aim to find hyperplanes in a high-dimensional data space that separates the objects belonging to different classes. A new object is classified by identifying its position relative to the built hyper planes [32];
- **Ensembles of classifiers** combine multiple classifiers to build a more robust classifier, usually using a voting mechanism. A very successful class of this type of algorithms is the boosting algorithms [33].

3. Methodology for electrical customers' characterization and classification

In order to formulate representative daily load profiles for different types of customers it is essential to assure good databases, which require a sufficient amount of recorded data and also a robust clustering model approach. Indeed, the quality of final decisions is directly related to the quality of data, which justifies the need for an initial data pre-processing step.

Basically, the methodology for electrical customers' characterization relies on the combination of unsupervised and supervised learning techniques. Clustering analysis is a class of unsupervised learning techniques, meaning that the data have no target attribute. It is intended to explore the data to find some intrinsic structures in them. In supervised algorithms, the classes are predetermined and these classes are then used to predict the values of the target attribute in future data instances.

The implemented methodology for electrical customers' characterization, based on a KDD procedure [34], is depicted in Fig. 1.

Finally, a classification model should learn rules that allow the attribution of a class to a new consumer. Shape indices could be derived from the representative daily load diagrams in order to obtain sense rules of satisfactory interpretation, and therefore, to express relevant information about the electricity consumers behaviour.

3.1. Data and feature selection

This first step includes the definition of the data sample which will be applied to the KDD process. Typically, the load profiling study is performed based on stored historical data and also additional commercial information concerning the electrical customers. In this stage, it is defined which type of customers will be chosen for analysis, i.e. low, medium or high voltage consumers.

This phase requires a good understanding and knowledge background of the study area, to choose and select judiciously the attributes related to the demand objectives, and only those, e.g. active and reactive power, voltage value, energy consumption, commercial code type, contracted power, time of day tariff period, peak power value, geographical location, etc. One other important task is the definition of the period of time that it is intended to analyze (season of the year, entire year), as well as the specification of the recorded interval cadence.

3.2. Data pre-processing

Real-world databases are highly susceptible to be inconsistent (discrepancies in data), incomplete (lacking attributes values) and/or noisy (containing errors or outlier values). The major obstacle to obtain knowledge is indeed poor data. There is the need to ensure that the knowledge discovery from the databases is in fact reliable.

By the fact that there are always problems with data, a pre-processing phase is required to detect and correct bad data. Typically, the pre-processing phase contains several sub-phases, namely, data cleaning, data integration, data selection and data transformation [10]. All these can be used within the proposed methodology. All the inconsistencies in the data are analyzed and outliers are removed based on the information of similar days. In the pre-processing step the lacking values are detected and replaced using ANN and also linear regression techniques [35].

Data are organized to represent the customer's electricity consumption by means of a typical daily load pattern. The data can also be distinguished according to different loading conditions into smaller datasets, e.g., for working days, Saturdays and Sundays and Holidays.

3.3. Determining typical load profiles

In this step, clustering algorithms are applied in order to perform load pattern grouping so that objects within a cluster have high similarity among them, and dissimilar to objects in other groups. The choice of a clustering algorithm depends on the existing database as well as the purpose of the task. Several clustering methods for cluster analysis are available to perform this task [10,36,37]. However, each clustering method may identify different clusters for the same data set. Thus, important questions are addressed, associated to the clustering procedure, namely, what is the best clustering method that produces the best data partition, and how many clusters k are presented in data. The evaluation of the clustering results is one of the most difficult issues in clustering analysis process. Many clustering validity indices have been defined to assess the effectiveness of the clustering process, namely to support the decision making concerning the choice of the best partition. Different validity measures have distinct criteria to evaluate a data partition. Thus, to have more confidence in the chosen partition, a variety of measures should be applied, and only then a decision should be taken considering the outputs of all validity indices.

To identify customer with the same consumption patterns from a certain database, N clustering algorithms are used to identify clusters contained in that database. Considering $k_{\min}$ and $k_{\max}$, respectively the minimum and maximum number of clusters for each clustering algorithm, $k_{\max} - k_{\min} + 1$ data partitions are generated by varying the number of clusters k between $k_{\min}$ and $k_{\max}$. Then, V clustering validity indices are used to evaluate the quality of all her data partitions produced by all the clustering algorithms.

$$D_p = (k_{\max} - k_{\min} + 1) \times \mathrm{N}$$

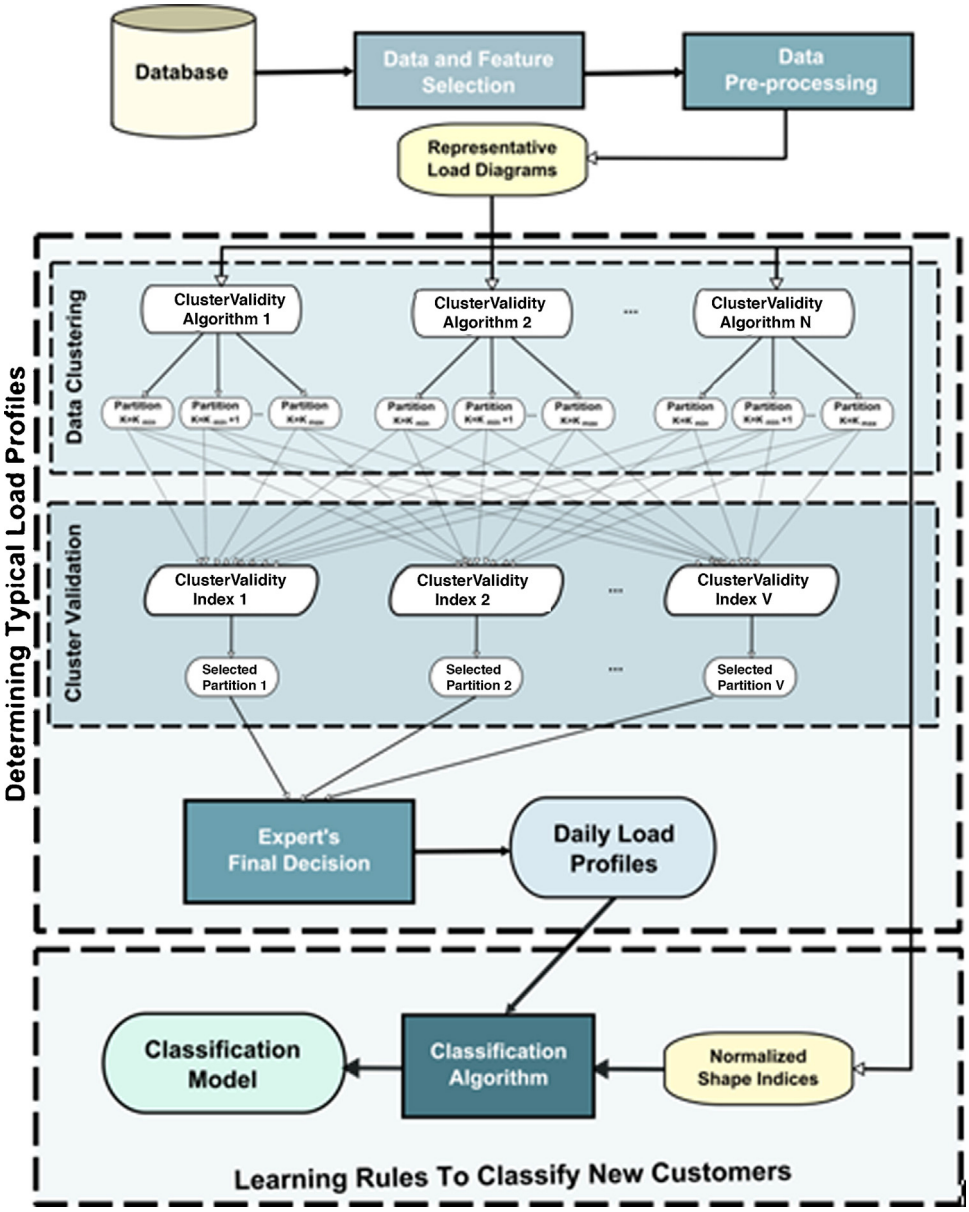


Fig. 1. Methodology to support electrical customers' characterization and classification.

The expert only has to decide, from a small subset of V data partitions, which partition is the best for his purpose (if there is no consensus between clustering validity indices in the identification of the best partition). The clustering results must be carefully analysed. One approach is to choose a partition result (for the same number of derived classes) based on the largest number of indices that points to the same result. In the case of index number equality the analyst will have to make a choice that might possibly not be the best. Follows that the involvement of experts is crucial in this evaluation phase of the result of clustering partition.

The typical daily load profiles are obtained by averaging the representative load profiles of electricity consumers in the same cluster.

3.4. Learning rules to classify new customers

In order to create a classification model to predict the class of a new consumer, a new representation of the daily load profiles should be created such that both the input data and the obtained classification rules are simple and intelligible. In this step it is recommended to use a rule-based modelling technique, such as a classification tree, since the resulting model is easily understandable.

For this purpose, the definition of specific indices related to daily load profiles is generally required. These indices can be defined taking into account commercial information, however, the use of commercial indices (i.e., contracted power, supply voltage level, commercial type of activity, etc.) to characterize the electricity customers' behaviour are totally unrelated to the load diagrams [4,35,38].

These indices could be directly derived from the daily load diagrams capturing relevant information about the electricity consumers' behaviour. The indices proposed in [38], and described in Table 1, include the load factor (f_1), the off-peak factor (f_2), the night impact coefficient (f_3), the lunch impact coefficient (f_4) and the modulation coefficient at off-peak hours (f_5), where P_{max} (P_{min})

Table 1
Normalized indices to characterize electricity customers' behaviour.

Parameter	Definition	Acquisition period	Reference
Daily P_{av}/P_{max}	$f_1 = P_{av,day}/P_{max,day}$	1 day	[38]
Daily P_{min}/P_{max}	$f_2 = P_{min,day}/P_{max,day}$	1 day	
Night Impact	$f_3 = 1/3P_{av,night}/P_{av,day}$	1 day (8 h night, from 11 p.m. to 6 a.m.)	
Lunch impact	$f_4 = 1/8P_{av,lunch}/P_{av,day}$	1 day (3 h lunch, from 12 a.m. to 15 p.m.)	
Daily P_{min}/P_{av}	$f_5 = P_{min,day}/P_{av,day}$	1 day	

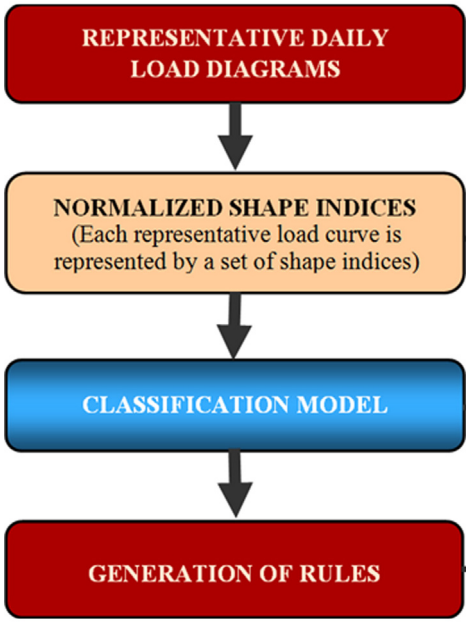


Fig. 2. Classification model architecture.

is the maximum (minimum) power demand of the representative day and P_{av} is the average power demand.

Fig. 2 illustrates the classification model architecture that was implemented.

The normalized shape indices, derived from the representative daily load diagrams, belong to the [0,1] range and are used as the attributes input in the classification model. A rule set is formed by the classification model to support the classification of new customers.

4. Case study

The case study uses 1.022 Medium Voltage (MV) customers which consumed power was recorded with a 15 min cadence in a period of one year, since September 1st 2010 until August 31st 2011. This sample was supplied by the Portuguese distribution company.

4.1. Data pre-processing

First, a data pre-processing step was conducted to analyze all the data and check for missing values. During this stage it was verified that on October 31st 2010 the files had recorded 25 h and only 23 h on March 27th 2011. This is explained by the daylight saving clock changes. For the first case, the records collected from 1:00 a.m. to 1:45 a.m. were removed. For the second case, the values for the missing hour, also from 1:00 a.m. to 1:45 a.m., were replaced by the power values of the previous day. After this step, all customers had the complete information of all the year and they were prepared to be used by the clustering algorithms.

After this data pre-processing step, a representative load curve was obtained by averaging the daily load diagrams of each customer. Therefore, each customer is represented by one typical load curve. However, these representative load profiles concern to the power consumption which means that the diagram curve is directly proportional to the amount of the electric energy consumption. As it is intended to compare the consumption pattern among customers, the power consumption was normalized to the [0,1] range, using the peak power of the each representative load diagram as normalization factor, maintaining this way the information related to the initial load profile shape. So, each customer is represented by a normalized representative daily load curve.

The normalized representative load curves of all customers were divided in three loading conditions, working days, Saturdays and Sundays and Holidays, because it is expected that during the working days the power consumption is different from that of the weekend and special days.

4.2. Load profiling

In this step it is intended to group the MV customers in classes following a similarity criterion. It is expected to group the load patterns on the basis of their distinguishing features. It was used the normalized representative daily load curve.

4.2.1. Typical load profile

The choice and selection of the clustering algorithm is crucial. In this case, it was based on previous authors' works [35,39,40] several clustering algorithms have been tested. Taking into account previous knowledge, in this case study the K-Means, N-Cut, PC-KMeans and MPC-KMeans algorithms were chosen.

To assess the data partitions, 8 validity indices presented in Section 2 were used: Normalized Hubert statistic index (*NH*), Dunn index (*D*), Davies–Bouldin index (*DB*), the SD validity index, Silhouette statistic (*S*), index *I*, Xie & Beni clustering validity index (*XB*) and Point symmetry index (*PS*).

The determination of the minimum and maximum number of clusters was performed in an iterative way. In the first place, in which it was imposed that the number of clusters K should vary between 2 and 20. That is, for each clustering algorithm was produced an index value within the defined range of K. From all the obtained data partitions, the one produced by K-Means algorithm with 3 clusters was selected as the best partition by 3 of the 8 used clustering validity indices. This partition is illustrated in Fig. 3.

However, by analyzing the plots of the partition, it was visually perceived that the clusters could be further divided. Therefore, a new clustering selection process was performed but considering only partitions with numbers of clusters from 4 to 17, because the minimum (and maximum) number of clusters should be higher (and less) than the minimum (maximum) number of clusters belonging to the set of partitions selected by the clustering validity indices.

In this case, once again, the best partition was produced by the K-Means algorithm with 4 clusters, supported by 5 validity indices (Fig. 4). A new clustering selection process was performed but only considering partitions with K varying from 5 to 14, in order to verify if the solution could still be improved. For this range, the best partition was produced by K-Means algorithm with 6 clusters, as shown in Fig. 5. The expert decided to stop the iterative process since the

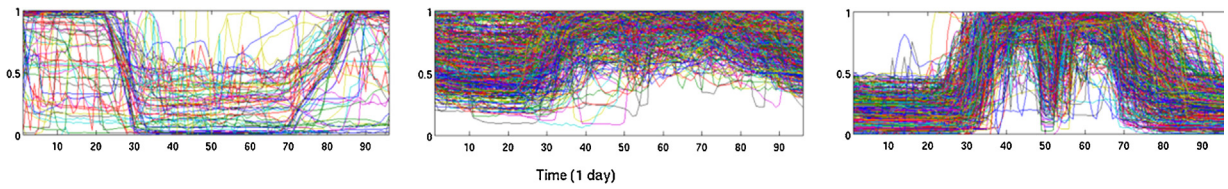


Fig. 3. K-Means algorithm results for k=3 clusters.

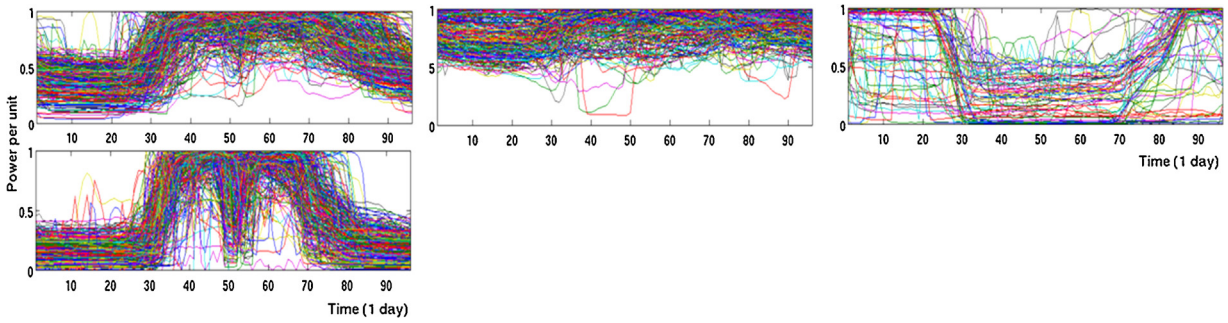


Fig. 4. K-Means algorithm results for k=4 clusters.

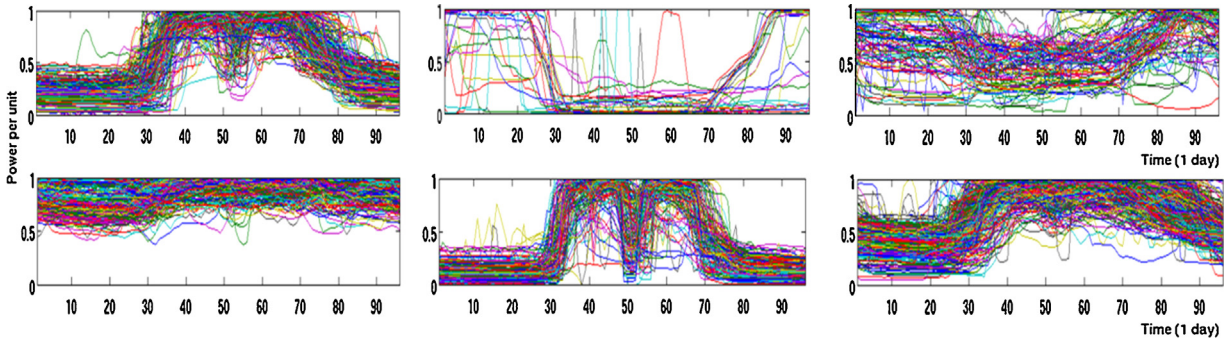


Fig. 5. K-Means algorithm results for k=6 clusters.

best partition obtained in the last iteration did not outperformed the quality of the best partition of the previous iteration.

This iterative process is summarized in Table 2, which shows the partition selected by each clustering validity index at each step. The partition selected by the majority of the clustering validity indices are shown in text bold.

After the choice of the best data partition, the normalized daily load diagrams have been used to produce the cluster representatives' curves. Thus, the representative diagram for each cluster, for weekends and working days, were obtained by averaging the representative load diagrams of the customers' assigned to the same cluster.

Figs. 6–8 show, respectively, the representative load diagram obtained for each cluster for working days, Saturdays and Sundays and holidays. Each curve represents the load profile of the corresponding customer class.

The results showed that the clustering module has well separated the customer population and representative load diagrams were created with distinct load shape.

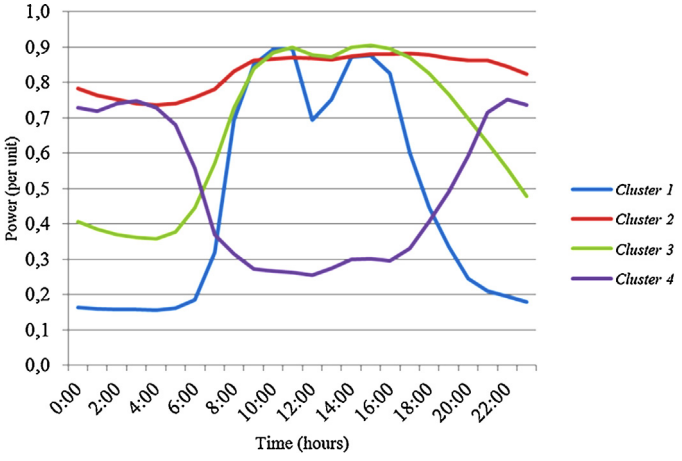


Fig. 6. Cluster representative load diagrams using normalized shape indices—working days.

4.2.2. Profile characterization

The indices defined in Table 1, which could be directly derived from the daily load diagrams, capturing relevant information about the electricity consumers' patterns consist of:

- Load factor (f_1)—Factor that represents how linear a load diagram is. If has a high value the diagram tends towards a straight line and if has a low value means that the consumption curve is not uniform over the time period under analysis, containing accentuated peaks and troughs;
- Off-peak factor (f_2)—Factor that emphasizes the impact of the curve line sags. High value means that there is low difference from the minimum to the highest value. Small value points to a

Table 2
Results of the partitions selected by the validity indices.

Cluster Range	Analysis of best result criteria	SD	PS	DB	XB	S	I	D	NH
2–20 Clusters	Cluster algorithm	KM	NC	KM	**KM**	**KM**	**KM**	MPCKM	PCKM
	Number of clusters	4	2	2	**3**	**3**	**3**	18	3
4–17 Clusters	Cluster algorithm	**KM**	**KM**	**KM**	PCKM	**KM**	**KM**	MPCKM	PCKM
	Number of clusters	**4**	**4**	**4**	5	**4**	**4**	15	5
5–14 Clusters	Cluster algorithm	KM	MPCKM	**KM**	PCKM	KM	**KM**	PCKM	PCKM
	Number of clusters	8	6	**6**	5	7	**6**	12	5

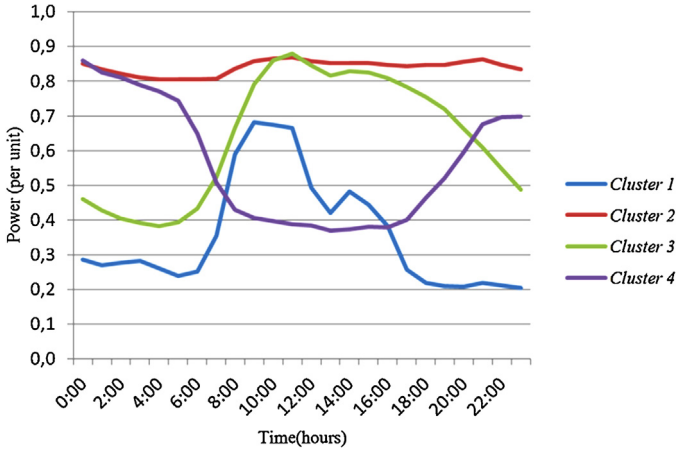


Fig. 7. Cluster representative load diagrams using normalized shape indices—Saturdays.

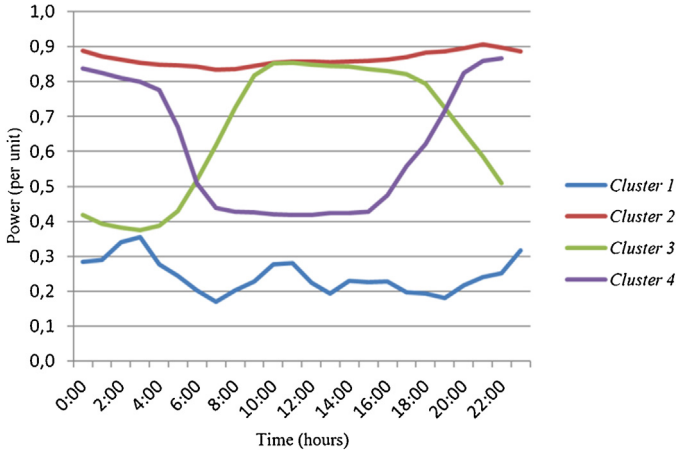


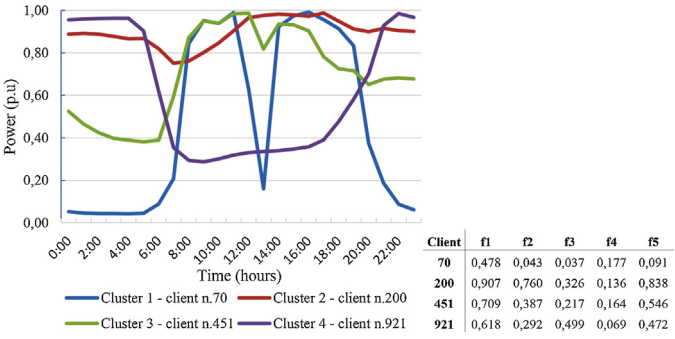
Fig. 8. Cluster representative load diagrams using normalized shape indices—Sundays and Holidays.

greater discrepancy between the minimum and maximum value observed in the load diagram;

- Night impact factor (f_3)—Factor that evaluates the impact of the electricity consumption during evening hours (11:00 p.m. to 6 a.m.). The higher the value of this index greater is the energy consumption in this period of time;
- Lunch impact factor (f_4)—Factor that evaluates the impact of the electricity consumption during lunch time hours (12:00 a.m. to 3 p.m.). The higher the value of this index greater is the energy consumption in this period of time;
- Modulation coefficient at off-peak hours (f_5)—Factor similar to f_2 but analyzing the discrepancy of the minimum value to the average consumption of the period in analysis. The smaller the value of this index greater is the discrepancy between the minimum and average consumption in the period of time in analysis.

Table 3
Classification indices results for the obtained clusters—working days.

Clusters	f_1	f_2	f_3	f_4	f_5
1	0.513	0.174	0.119	0.217	0.339
2	0.938	0.834	0.307	0.132	0.889
3	0.732	0.395	0.200	0.168	0.539
4	0.656	0.339	0.476	0.072	0.517



Client	f1	f2	f3	f4	f5
70	0,478	0,043	0,037	0,177	0,091
200	0,907	0,760	0,326	0,136	0,838
451	0,709	0,387	0,217	0,164	0,546
921	0,618	0,292	0,499	0,069	0,472

Fig. 9. Normalized shape indices of a cluster consumer.

Table 3 reports the results of the indices defined in Table 1 applied to the obtained typical load profile, in this example for the representative load diagrams for weekdays (Fig. 6). This indices were used as input attributes in the classification model in order to achieve intelligible rules.

Analyzing Fig. 6 and also the results of Table 3 it is clear the correlation between the load diagram patterns and the indices results. Therefore, the highest value of f_1 belongs to cluster 2 which is the smoother curve among others. The cluster that has the lowest f_1 factor value is cluster 1. In this case, it is obvious the difference between the highest and the lowest consumption value. The cluster 2 pattern presents the lowest difference between the lowest and highest consumption value, consequently f_2 and f_5 presents the highest value. In opposite, cluster 1 presents the highest difference between lowest and highest consumption value, as a result f_2 and f_5 presents the lowest value. The consumers that belong to cluster 4 have high consumption during evening hours, therefore the night impact factor (f_3) presents the highest value. The highest value of lunch impact factor (f_4) belong to cluster 1 as a result to have an intensive consumption during the specified lunch time.

Fig. 9 depicts a customer from each obtained cluster and also the value of the normalized shape indices. The obtained indices values describes the load diagram pattern of each selected MV consumer.

4.3. Rules definition for customer classification

In order to obtain more relevant information to describe the consumption patterns of each cluster population a rule-based modelling technique has been used, in this case the C5.0 classification algorithm. This algorithm, which is a decision tree, produces rules that are easy to understand and have an intelligible interpretation.

Table 4
Rule set of the classification model – working days.

If $f_1 \leq 0.681 \wedge f_3 \leq 0.277 \wedge f_4 \leq 0.154 \wedge f_1 \leq 0.461$	**cluster – 4**
If $f_1 \leq 0.681 \wedge f_3 \leq 0.277 \wedge f_4 \leq 0.154 \wedge f_1 > 0.461 \wedge f_1 \leq 0.640 \wedge f_1 \leq 0.579$	**cluster – 1**
If $f_1 \leq 0.681 \wedge f_3 \leq 0.277 \wedge f_4 \leq 0.154 \wedge f_1 > 0.461 \wedge f_1 \leq 0.640 \wedge f_1 > 0.579 \wedge f_5 \leq 0.556$	**cluster – 3**
	then **cluster – 1**
If $f_1 \leq 0.681 \wedge f_3 \leq 0.277 \wedge f_4 \leq 0.154 \wedge f_1 > 0.461 \wedge f_1 > 0.640$	**cluster – 3**
If $f_1 \leq 0.681 \wedge f_3 \leq 0.277 \wedge f_4 > 0.154 \wedge f_1 \leq 0.364 \wedge f_1 \leq 0.306$	**cluster – 4**
	then **cluster –1**
If $f_1 \leq 0.681 \wedge f_3 \leq 0.277 \wedge f_4 > 0.154 \wedge f_1 > 0.364$	**cluster – 1**
If $f_1 \leq 0.681 \wedge f_3 > 0.277 \wedge f_3 \leq 0.352 \wedge f_2 \leq 0.244$	**cluster – 4**
If $f_1 \leq 0.681 \wedge f_3 > 0.277 \wedge f_3 \leq 0.352 \wedge f_2 > 0.244 \wedge f_4 \leq 0.115$	then **cluster – 2**
	then **cluster – 3**
If $f_1 \leq 0.681 \wedge f_3 > 0.277 \wedge f_3 > 0.352$	**cluster – 2**
If $f_1 > 0.681 \wedge f_4 \leq 0.097$	**cluster – 2**
If $f_1 > 0.681 \wedge f_4 > 0.097 \wedge f_4 \leq 0.167$	**cluster – 3**
If $f_1 > 0.681 \wedge f_4 > 0.097 \wedge f_4 > 0.167 \wedge f_3 \leq 0.204 \wedge f_1 \leq 0.706$	**cluster – 1**
	then **cluster – 3**
If $f_1 > 0.681 \wedge f_4 > 0.097 \wedge f_4 > 0.167 \wedge f_3 > 0.204$	**cluster – 3**

Table 5
Rule set classification for the real MV customer.

Parameter	$f_1 = 0.567$	$f_2 = 0.247$	$f_3 = 0.161$	$f_4 = 0.215$	$f_5 = 0.435$	**Cluster 1**
Classification rule	If $f_1 \leq 0.681 \wedge f_3 \leq 0.277 \wedge f_4 \leq 0.154 \wedge f_1 > 0.461 \wedge f_1 \leq 0.640 \wedge f_1 \leq 0.579$					

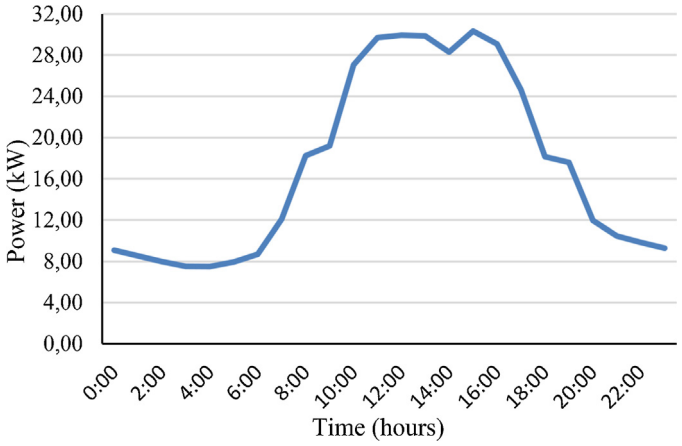


Fig. 10. Typical chronological load profile of a real MV customer—Weekdays.

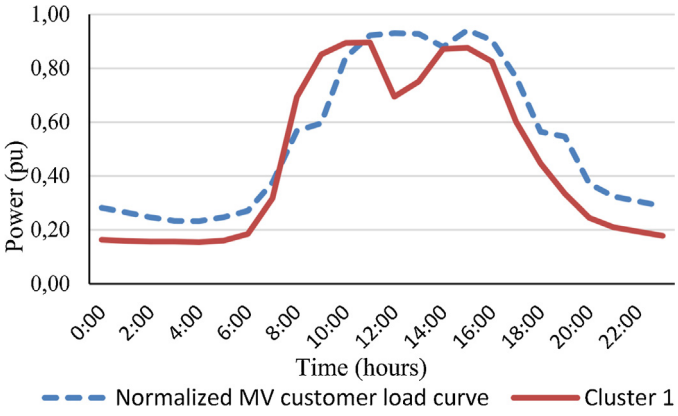


Fig. 11. Normalized load profile of the real MV customer and cluster 1—Weekdays.

Normalized shape indices were used as attributes in the classification model. The indices vector f is formed by the indices presented in Table 1.

The working days and weekend data sets were split to form the training and the testing data sets by using 2/3 of the data for training, and the remaining for testing.

Table 4 presents a rule set example obtained by the C5.0 algorithm from working days data set. The obtained rules are simple and easy to understand.

The classification model used all the available attributes, selecting for each rule merely the attributes that provided larger information gain. The model testing accuracy was 94.83% for working days and 95.10% for weekend days, which means that these simple rules are highly accurate.

4.3.1. Classification of a real MV customer

In this case study it is proposed to classify a real MV customer taking into consideration the analysis of its electricity consumption. The power consumption data of this MV customer was supplied by the Portuguese distribution company—EDP-Distribuição. Following the proposed electricity customer characterization methodology, the data concerning electric energy consumption were subjected to the several steps identified in Section 3: data and feature selection; data pre-processing and determining typical load profiles. As result, Fig. 10 illustrates the typical chronological load curve, in this example referring to the weekdays (Monday to Friday).

Table 5 shows the indices curve pattern results by application of the formulation presented in Table 1. Taking into account its chronological energy usage, this electricity customer was classified in the cluster 1.

Fig. 11 depicts both load curves, normalized load curve of the real MV customer and cluster 1. One can confirm the pattern similarity of both curves and verify that the customer was well categorized.

In an electricity market environment, each supplier company desires to identify his customers' electricity behaviour accurately, to provide them with satisfactory services at a better low cost. In this scope, the categorisation of electricity customers revels as a necessary and important stage. Concomitantly, each consumer wishes to know his own electricity consumption behaviour, in order to apply energy efficiency measures successfully or to select the most appropriate tariff structure.

5. Concluding remarks

This paper presents a methodology for the characterization of medium voltage electric consumers. Different clustering techniques have been assessed in order to identify similar load patterns

among electricity customers. A real database was used containing historical consumption data related to 1.022 MV customers, during an entire year. In the implemented methodology for electrical customers' characterization, based on the knowledge discovery in databases process, all steps of this procedure were outlined, namely the database definition, pre-processing data, data mining applications, approach methodology to the decision of the best partition and suitable number of clusters, classification model, and finally, the interpretation of the discovered knowledge.

Several clustering algorithms were used to obtain the typical load profiles and 8 clustering validity indices were applied to help identifying the best one. Four distinct typical load profiles were identified where each curve was clearly different from the others.

A classification model was used to enable the classification of new consumers using a set of normalized shape indices as features. A decision tree was chosen for this task as it returns rules that helps explaining the electricity consumption customers' behavior. The classification algorithm presents a good overall accuracy for both working days and weekend loading conditions.

The distribution companies, as well as the consumers, can take many advantages from the knowledge of the typical load profile and this knowledge can improve the electric power supplier-consumers agreements. Load profiling may enable the establishment of business contracts between distributers and suppliers in the liberalized market.

Acknowledgements

This work is supported by FEDER Funds through COMPETE program and by National Funds through FCT under the projects FCOMP-01-0124-FEDER: PEst-OE/EEI/UI0760/2015, and PTDC/SEN-ENR/122174/2010, and by the GID-MicroRede, project no. 34086, co-funded by COMPETE under FEDER via QREN Programme. The present work is also developed under the EUREKA—ITEA2 Project SEAS with project number 12004.

The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under project ELECON, REA grant agreement no. 318912.

References

[1] C.-S. Chen, et al., Implementation of the load survey system in Taipower, in: IEEE Transmission and Distribution Conference vol. 1, 11–16 April 1999, 1999, pp. 300–304.
[2] Runming Yao, Koen Steemers, A method of formulating energy load profile for domestic buildings in the UK, Energy Build. 37 (6) (2005) 663–671.
[3] J. Widén, M. Lundh, I. Vassileva, E. Dahlquist, K. Ellegård, E. Wäckelgård, Constructing load profiles for household electricity and hot water from time-use data—modelling approach and validation, Energy Build. 41 (7) (2009) 753–768.
[4] V. Figueiredo, F. Rodrigues, Z. Vale, B. Gouveia, An electric energy characterization framework based on data mining techniques, IEEE Trans. Power Syst. Vol. 20 (May (N.2)) (2005) 596–602.
[5] Gianfranco Chicco, Overview and performance assessment of the clustering methods for electrical load pattern grouping, Energy 42 (June (1)) (2012) 68–80.
[6] Young-Il Kim, Jin-Ho Shin, Jae-Ju Song, Il-Kwan Yang, Customer clustering and TDLP (typical daily load profile) generation using the clustering algorithm, in: Transmission & Distribution Conference & Exposition: Asia and Pacific, 26–30 October 2009, 2009, pp. 1–4.
[7] The European Parliament and the Council of European Union, Directive 2009/72/EC of the European Parliament and of the Council of 13 July 2009, The European Parliament and the Council of European Union, 2009 (July).
[8] D. Gerbec, S. Gasperic, I. Smon, F. Gubina, Allocation of the load profiles to consumers using probabilistic neural networks, IEEE Trans. Power Syst. 20 (May (no. 2)) (2005) 548–555.
[9] I.P. Panapakidis, M.C. Alexiadis, G.K. Papagiannis, Electricity customer characterization based on different representative load curves, in: Ninth International Conference on the European Energy Market, 10–12 May, 2012, pp. 1–8.
[10] J. Han, M. Kamber, Data Mining: Concepts and Techniques The Morgan Kaufmann Series in Data Management Systems, San Francisco, 2006.
[11] K.Jain. Anil, Data Clustering: 50 years beyond K-Means, Pattern Recognit. Lett. 31 (June (8)) (2010) 651–666.
[12] C.C. Aggarwal, C.K. Reddy, Data Clustering: Algorithms and Applications Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, CRC Press, Boca Raton, FL, 2013.
[13] J.B. MacQueen, Some Methods for classification and Analysis of Multivariate Observations, in: Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, 1967, pp. 281–297.
[14] S. Basu, A. Banerjee, R.J. Mooney, Active semi-supervision for pairwise constrained clustering, in: Proceedings of Fourth SIAM International Conference on Data Mining, 2004.
[15] S. Basu, Mikhail Bilenko, R.J. Mooney, Comparing and unifying search-based and similarity-based approaches to semi-supervised clustering, in: Proceedings of the ICML-2003 Workshop on the Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining Systems, Washington DC, 2003 (August), pp. 42–49.
[16] K.A. Jain, C.R. Dubes, Algorithms for Clustering Data, in: Prentice-Hall Advanced Reference Series, Prentice-Hall, Inc., Upper Saddle River, NJ, 1988.
[17] M. Ester, H.P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proc. Second Int. Conf. on Knowledge Discovery and Data Mining, AAAI Press, Portland, OR, 1996, pp. 226–231.
[18] P. Grabusts, A. Borisov, Using grid-clustering methods in data classification, in: Proceedings of the International Conference on Parallel Computing in Electrical Engineering, 2002.
[19] Jianbo Shi, Jitendra Malik, Normalized cuts and image segmentation, in: IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 2000.
[20] C. Fraley, A.E. Raftery, Model-based clustering, discriminant analysis, and density estimation, J. Am. Stat. Assoc. 97 (458) (2002) 611–631.
[21] Halkidi Y. Batistakis, M. Vazirgiannis, Clustering algorithms and validity measures, in: Tutorial Paper in the Proceedings of the SSDBM, Conference, 2001.
[22] P. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65.
[23] L. Hubert, P. Arabie, Comparing partitions, J. Classif. 2 (1985) 193–218.
[24] J.C. Dunn, Well separated clusters and optimal fuzzy partitions, J. Cybern. 4 (1974) 95–104.
[25] D.L. Davies, D.W. Bouldin, A cluster separation measure, IEEE Trans. Pattern Anal. Mach. Intell. 1 (2) (1979).
[26] U. Maulik, S. Bandyopadhyay, Performance evaluation of some clustering algorithms and validity indices, IEEE Trans. Pattern Anal. Mach. Intell. 24 (12) (2002) 1650–1654.
[27] C. Chou, M. Su, E. Lai, Symmetry as a new measure for cluster validity, in: Second WSEAS International Conference on Scientific Computation and Soft Computing, 2002, pp. 209–213.
[28] X. Xie, G. Beni, A validity measure for fuzzy clustering, IEEE Trans. Pattern Anal. Mach. Intell. 13 (1991) 841–847.
[29] Geoffrey J. Mclachlan, Discriminant Analysis and Statistical Pattern Recognition Wiley Series in Probability and Statistics, Wiley-Interscience, New York, 2004 August.
[30] Nilima Patil, Rekha Lathi, Vidya Chitre, Comparison of C5.0 & CART classification algorithms using pruning technique, Int. J. Eng. 1.4 (2012) 1–5.
[31] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning internal representations by error propagation, in: E. David, Rumelhart, L. James, CORPORATE. McClelland, P.D.P. Research Group (Eds.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1, MIT Press, Cambridge, MA, USA, 1986, pp. 318–362.
[32] Nello Cristianini, John Shawe-Taylor, An Introduction to Support Vector Machines: And Other Kernel-Based Learning Methods, Cambridge University Press, New York, NY, USA, 1999.
[33] Yoav Freund, Robert E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting", J. Comput. Syst. Sci. 55 (1997) 119–139.
[34] U.G. Fayyad, G. Piatetsky-Shapiro, P.J. Smith, R. Uthurasamy, From data mining to knowledge discovery: an overview, in: Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, Cambridge, 1996, pp. 1–34.
[35] S. Ramos, Z. Vale, J. Santana, J. Duarte, Data mining contributions to characterize MV consumers and to improve the suppliers-consumers settlements, in: Proc. 2007 IEEE/PES General Meeting, 24–28 June, Tampa, Florida, USA, 2007.
[36] E. Carpaneto, G. Chicco, R. Napoli, M. Scutariu, Electricity customer classification using frequency—domain load pattern data, Int. J. Electr. Power Energy Syst. Volume 28 (1) (2006) 13–20.
[37] A.M.S. Ferreira, C.A.M.T. Cavalcante, C.H.O. Fontes, J.E.S. Marambio, A new method for pattern recognition in load profiles to support decision-making in the management of the electric sector, Int. J. Electr. Power Energy Syst. 53 (December) (2013) 824–831.
[38] G. Chicco, R. Napoli, P. Postolache, M. Scutariu, C. Toader, Customer characterization options for improving the tariff offer, IEEE Trans. Power Systems 18 (February) (2003) 381–387.
[39] F.J. Duarte, A. Fred, F. Rodrigues, J.M. Duarte, S. Ramos, Z. Vale, Determination of electricity consumers' load profiles via weighted evidence accumulation clustering using subsampling, in: Third IASTED Asian Conference on Power and Energy Systems, Thailand, 2007 April.
[40] S. Ramos, J.M. Duarte, F.J. Duarte, Z. Vale, P. Faria, A data mining framework for electric load profiling, in: IEEE PES Conference on Innovative Smart Grid Technologies – ISGT Latin America 2013, São Paulo, Brazil, 2013 April.