Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2020) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

14th International Symposium «Intelligent Systems», INTELS'20, 14-16 December 2020, Moscow, Russia

Data mining techniques for electricity customer characterization

Sérgio Ramos[a]*, João Soares[a], Samuel S. Cembranel[a, b], Inês Tavares[a], Z. Foroozandeh[a], Zita Vale[a], Rubipiara Fernandes[b]

[a]*GECAD, Polytechnic of Porto, Rua Dr. Bernardino de Almeida, 431, 4249-015, Porto, Portugal*
[b]*Instituto Federal de Santa Catarina (IFSC), Av. Mauro Ramos, 950 - Centro, Florianópolis, 88020-300, Brazil*

Abstract

The liberalization of electricity markets has been resulted in the emergence of new players, increasing the competitiveness in the electricity sector, aiming to provide better services and better prices. The knowledge of energy consumers' profile has been an important tool to help players to make decisions in the electrical sectors. In this paper, a characterization model of typical load profiles for Low Voltage (LV) customers is proposed and evaluated. The identification of consumption patterns is based on clustering analysis. The clustering methodology is based on seven clustering algorithms (partitional and hierarchical). Also, five clustering validity indices are used to identify the best data partition. With the knowledge obtained in clustering analysis, a classification model is implemented in order to classify new customers according to their consumption data. The classification model is used to select the correct class for each customer. To simplify the classification model, each load curve is represented by three indices which represent the load curves shape. The methodology used in this work demonstrates to be an effective tool and can be used in most diverse sectors, highlighting the use of knowledge in the optimization of the energy contracting for low voltage consumers. The energy consumption data can be constantly updated to improve the model precision, as well to better represent consumers and their consumption habits.

Keywords: Knowledge Discovery in Databases; Data Mining; Clustering; Classification; Typical Load Profiles.

1. Introduction

In the last decades, electricity markets are facing some changes. The great change in these markets is the liberalization of sectors such as energy generation, transmission, distribution, and commercialization. The liberalization of energy contracting for customers, in most countries, has been occurred in phases, by starting to include

* Corresponding author. Tel.: +351 228 340 500; fax: +351 225 020 772
E-mail address: scr@isep.ipp.pt

Peer-review under responsibility of the scientific committee of the 14th International Symposium «Intelligent Systems».

large customers (higher voltage levels) for then reducing the energy consumption level until reach the lower level (low voltage).

Recently, the Brazilian energy market is facing a great change in economic dispatch and energy pricing policy. Until December 2019, the energy pricing and economic dispatch were calculated weekly and they were based on three load levels: peak, off-peak, and baseload. Since January 2020 the power plants dispatch has been calculated for each day, for the day ahead, and discretized in 48 intervals (one for each half-hour of the day), but the energy price continues to be calculated weekly. According to the Brazilian National Electricity Agency (ANEEL), it tends to change for 2021 and the economic dispatch will be calculated daily with 48 intervals. Also, the price will be calculated daily with 24 intervals (one for each hour) [1].

Figure 1 presents the comparison between hourly prices and weekly prices in Brazil. It is possible to see a margin of monetary gain, by consuming more energy in the first seven hours of the day and save energy during the rest of the day. Is can be observed that the knowledge of a customer consumption pattern can help to make decisions in order to improve energy consumption and it is also important to reduce energy costs.

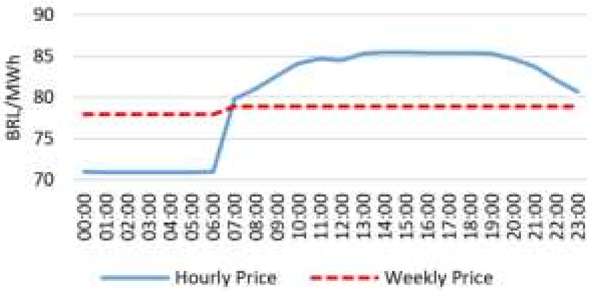

Fig. 1. Comparison between hourly prices and weekly prices in Brazil.

The knowledge about consumption patterns can be also used for energy efficiency programs, electricity price policies, load forecasting, demand-side management (DSM), and customer classification.

To identify the consumption patterns in a dataset is necessary to use Data Mining (DM) techniques to treat the data and find patterns, for then evaluate the knowledge. One of the main DM techniques used to find patterns is clustering. The clustering analysis is used when there is no previous knowledge about the dataset. Otherwise, when there is a previous knowledge about the dataset, classification techniques can be used to obtain the patterns. During the last decades, a lot of effort was been done in an attempt to use DM techniques in electricity markets with a focus to improve energy services.

To better use energy, reference [2] proposes a system to improve the energy consumption in a building. As is shown in the study, the system is capable to reduce 21% of the energy consumption. By using clustering algorithms, reference [3] analyses the impact of refrigerators replacement for low voltage customers, the tool developed helps to take decisions about the energy efficiency and management. The reference [4] develops an algorithm capable to identify energy consumption patterns and classify customers. The study also emphasizes that the typical load profiles (TLP) extracted from the dataset can help in energy systems management by improving the energy consumption with Demand Response (DR).

One of the most important tools in electrical power systems is load forecasting. The knowledge of the load pattern can be used to maintain voltage and frequency levels, it can improve grid utilization by promoting DR programs, reduce energy costs, and decrease peak load. DM techniques contribute to this analysis by treating datasets and extracting relevant patterns to analyze with forecast algorithms [5–10].

In modern electricity sectors, DR is emerging as a tool to improve the utilization of the grid. The pattern recognition by using DM techniques can support DR programs. The knowledge of energy consumption patterns can help to estimate the best period to reduce energy consumption, improving energy utilization, and decreasing costs [11].

The objective of energy tariff is to reflect the cost of energy production, transmission, and distribution. In [12] tariff structures are proposed based on clustering analysis. The purpose of the study is to find a tariff that can better represent the power grid utilization, by increasing energy prices when the energy consumption is high and reducing the price when the consumption is low. It gives the economic sign for customers to reduce energy consumption when the price is high, trying to improve the power grid utilization. Another proposal for the utilization of DM techniques is described in [13], by proposing tariff options based on customers patterns. In [14] it was identified that in some cases, energy contractual parameters are not related to the consumption patterns. In this study, it is also proposed a new tariff option to reduce the energy peak in the power network.

After the analysis of some DM applications, it is important to evaluate and decide the best techniques to extract load patterns from databases. With the intent to identify the best algorithm to clustering load diagrams, the reference [15] evaluates nine clustering algorithms, including agglomerative hierarchical clustering, K-means, Diana, Partitioning around Medoids, Clara, Fanny, Self-Organizing Maps (SOM), Model-based clustering and SOTA. To evaluate the partitions and the best clustering algorithms, three validity indices were used: Connectivity, Dunn, and Silhouette Index. The best results are from the PAM algorithm with two clusters, Diana with nine clusters to Dunn Index, and SOM to Silhouette Index with three clusters.

Three clustering algorithms are evaluated in reference [16], to identify the best clustering algorithm and the best partitions by using the Davies-Bouldin Index. The SOM algorithm proved to be the best, compared with K-means and K-medoids algorithm. The work developed by the reference [17] compares several clustering algorithms by using six clustering validity measures. The authors have developed a K-means method that shows better results for most of the metrics.

In [18] a framework to characterize medium-voltage consumers is proposed. The work uses clustering to extract load patterns and uses eight validity indices to evaluate the best clustering algorithm partition, as well as the optimal number of clusters. The K-means algorithm shows the best results for most of the metrics. After the clustering phase, the knowledge is used to train a classification model to classify new customers in one of the obtained clusters. The reference [19] proposes a characterization model divided into two parts: the first part is based on clustering, where the load patterns are extracted from the dataset by using SOM and K-means algorithms; the second part establishes a classification model, based on the C5.0 algorithm, which is used to classify electricity customers.

In [20] several classification algorithms were compared and the algorithm C5.0 (decision tree) shows high accuracy to classify electricity customers. The same reference uses fifteen load shape indices to find a better representation of the load diagrams.

This paper proposes a characterization framework based on clustering and classification to achieve the characterization of low voltage electricity customers. After this introduction, Section 2 presents some concepts of the Knowledge Discovery in Databases (KDD) process and the methodology proposed in this paper. Section 3 describes the clustering concepts and the algorithms used in this proposed work. Section 4 shows the clustering validity assessment used in this paper, to identify the best clustering algorithm as well as the best number of partitions. Section 5 presents the classification algorithm used to classify new electricity customers. Section 6 presents the case study and the obtained results. Finally, Section 7 draws conclusions and future works.

2. Knowledge Discovery Process in Databases

The 2. Knowledge Discovery Process in Databases (KDD) process is a merge of techniques to treat and discover patterns contained in data. In this paper, the proposed framework is based on the KDD process, illustrated in Figure 2. This framework includes several steps, starting with data selection, following by data pre-processing phase, the DM techniques stage, and finally, the knowledge evaluation.

2.1. Data Selection

The first step in the KDD process is to define the problem. After that, it is possible to select the data to apply algorithms to solve the problem. For example, to explain electricity consumption patterns in a certain area is required to identify the most relevant customer, such as, residential, commercial, or industrial consumers. Another possible selection of customers is by their contractual parameters, such as voltage levels, since then there is no sense in a comparison with energy consumption patterns between low and high voltage customers.

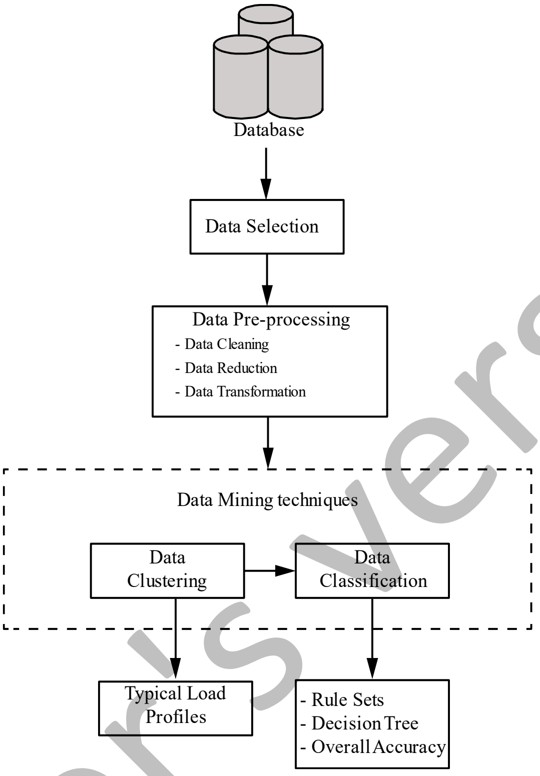

Fig. 2. Customers' characterization proposed methodology.

2.2. Data pre-processing

The energy consumption data acquisition can involve some errors and inconsistencies, such as equipment precision, communication failure between equipment and the dataset, data conversion, and storage. These problems can cause outliers, duplicate values, and lack of values. To minimize these errors, a pre-processing phase is applied to the dataset. Firstly, the data need to be converted to a common unit, for example kWh or MWh, to avoid wrong interpretation of the data. After that, missing values are filled out using prediction techniques, such as regression, the use of a global value, or an average of neighbouring points and neural networks. Outliers can be removed by using clustering techniques or statistical analysis and by avoiding them it is possible to have a better interpretation of the data and avoid wrong tendencies. The last two pre-processing techniques are data reduction and data transformation. Data reduction is required to reduce the amount of the data, and to improve DM analysis in order to find the relevant patterns contained in the dataset. In this case, load diagrams can be simplified according to the weekdays, weekends, holidays, or the seasons of the year. The data transformation is essential in the analysis, once it is intended to compare patterns and not amount of energy consumption, so the data need to be transformed to compare the patterns.

To transform the data, normalizations such as min-max can be applied. Other normalizations that can be applied to the data is the division of all the data by the max value to have a similar scale of the min-max normalization, but in this conversion, the lower value is not zero (unless there are values equal zero in the scale), but something near from zero. Another normalization is the division of the data by the average of the dataset, obtaining a per-unit value of the series. This phase is the most important in the KDD process and it is probably the most time and effort expanding phase. If this phase is not performed correctly, it can lead to the discovery of wrong patterns, resulting in errors in the interpretation of the problem.

2.3. Data Mining step

After select and treat the data, the DM algorithms can be applied to discover patterns in the dataset. Different techniques can be applied to extract the relevant information. The most used techniques are clustering, classification, and association rules. In this analysis, clustering algorithms are going to be used to identify the typical load profiles, and next, a classification algorithm is going to be applied in an attempt to classify new customers based on their historical data and the knowledge obtained by the clustering algorithms.

2.4. Knowledge

At the end of the KDD process, the knowledge is achieved and in this phase the patterns obtained are evaluated to solve or explain the initial problem. The obtained knowledge could be interesting if can added new knowledge from the existing one.

3. Clustering

Clustering is the formalization of mathematical techniques to identify patterns contained in databases [21]. A cluster can be considered as a class of the dataset and the data inside one cluster are characterized to be more similar to each other than the data in the other clusters. During the years, some algorithms were developed, the oldest and the main used techniques are the partitional and hierarchical. Other algorithms were developed in an attempt to find better results than the traditional techniques and they are based on density, fuzzy techniques, artificial neural networks, and evolutionary methods.

In this section, we will present the algorithms used in this work, these algorithms are based on partitional techniques and hierarchical techniques.

3.1. K-means algorithm

The K-means algorithm [22] is a partitional technique. The objective of this algorithm is to divide the dataset into a pre-established number of partitions (K). The K-means is probably the most known technique used in clustering analysis because it is a simple algorithm and shows good results compared with other clustering algorithms. The K-means can be used as a benchmark to test other new clustering methodologies.

One efficient version of this algorithm was presented by [22], and it can be summarized in five steps:

- 1. Select the number of clusters (parameter K);
- 2. Initialize the K cluster centers. There are two ways to do it, they can be chosen randomly or if the dataset is known, it is possible to choose the initial centers to improve the results;
- 3. Find the closest cluster center for each point. The distance from the cluster center to each point is calculated by the Euclidean distance;
- 4. Update the cluster center by calculating the average of the points belonging to each cluster.
- 5. The K-means will repeat the steps 2, 3, and 4 until the convergence criterion will be attended, or the cluster centers do not change.

In this algorithm different initializations (initial data points) can lead to different results because the algorithm tries to find a local optimum and not a global optimum. The use of K-means algorithm to clustering load profiles problems can be seen in [15], [17], [23–25].

3.2. K-medoid algorithm

The K-medoid algorithm [26] is a partitional algorithm, such as K-means. There are three main differences between these two algorithms. The K-means tries to minimize the total squared error and the K-medoid tries to minimize the total dissimilarity between the objects. The K-means uses the Euclidean distance to evaluate the distance between two data points and the K-medoid uses the Manhattan distance. Another difference between these two algorithms is that the K-means chooses the mean as the cluster center and the K-medoid chooses one point to be the cluster center (the medoid).

3.3. G-means algorithm

The G-means algorithm was proposed in reference [27]. This algorithm is an automatic clustering method. The great advantage of this algorithm is that unlike the K-means, this algorithm does not need a pre-established parameter (the number of clusters) to divide the dataset. The G-means algorithm is based on a test to see if the data assigned to each cluster center follows a Gaussian distribution. The algorithm proceeds according to the following steps [27]:

- 1. Select initial centers.
- 2. Runs K-means to cluster the dataset;
- 3. Test the data to see if the data of each cluster follows a Gaussian distribution;
- 4. If the data assigned to each cluster center follow a Gaussian distribution keep the center, on the contrary, replace for two new centers;
- 5. Repeat steps 2 to 4 until all data assigned to each cluster look Gaussian.

3.4. Hierarchical algorithms

The hierarchical techniques are divided into two categories, agglomerative and divisive. The agglomerative techniques start with all data as a cluster and in each iteration the clusters are merged in a new cluster by their proximity. At the end of the iterative process, all data are assigned to one cluster. The divisive analysis starts with all data in a cluster, and in each iteration the cluster is divided into smaller clusters until each data is one cluster [21].

One advantage of hierarchical techniques, compared to partitional techniques, is that these techniques do not need a pre-established parameter (number of partitions) to partition a dataset.

There are some ways to evaluate the distance between two clusters. Considering three clusters, a, b, and c, the clusters a and b are merged into a new cluster called n. The reference [28] proposes a general equation (1) to calculate the dissimilarity between the clusters.

$$(c,n) = \alpha_a d(c,a) + \alpha_b d(c,b) + \beta d(a,b) + \gamma|d(c,a) - d(c,b)| \tag{1}$$

By changing the values of the coefficients α, β and γ is possible to find the equation of the well-known hierarchical agglomerative algorithms in Table I of the next section [29].

4. Clustering validity indices

The choice of the right number of clusters is not an easy process without mathematical mechanisms, mainly when the dataset has a high volume of data. To help in clustering analysis and decision making, mathematical mechanisms are used to identify the best number of partitions and the quality of the partitions of the database.

Different clustering validity indices have been proposed to identify the effectiveness of clustering methods and identify the best number of partitions of a certain dataset. In this work, it was proposed to use 5 clustering validity indices in an attempt to identify the best number of partitions and also to evaluate the clustering algorithms performance.

The Mean Index Adequacy (MIA) evaluates the distance between the data associated to each cluster to the cluster center. This metric reflects the compaction of the clusters, so a lower level of this index is better. The Clustering Dispersion Indicator (CDI) is the relation of the distance between the data in each cluster (measure of compaction) and the distance between the clusters' centers (well separated clusters). So, a lower level of CDI indicates compacted and well separated clusters. The mathematical formulation of these two indices can be seen in the reference [13].

Table 1. Parameters for hierarchical clustering methods

Method	α_a	α_b	β	γ
Single link	1/2	1/2	0	-1/2
Complete link	1/2	1/2	0	1/2
Unweighted average (UPGMA)	$\frac{n_a}{n_a+n_b}$	$\frac{n_b}{n_a+n_b}$	0	0
Weighted average (WPGMA)	1/2	1/2	0	0
Unweighted centroid (UPGMC)	$\frac{n_a}{n_a+n_b}$	$\frac{n_b}{n_a+n_b}$	$\frac{-n_a n_b}{(n_a+n_b)^2}$	0
Weighted centroid (WPGMC)	1/2	1/2	-1/4	0
Minimum Variance (Ward)	$\frac{n_a+n_n}{n_a+n_b+n_n}$	$\frac{n_a+n_n}{n_a+n_b+n_n}$	$\frac{-n_n}{n_a+n_b+n_n}$	0
Flexible (Lance and Williams)	$\frac{1-\beta}{2}$	$\frac{1-\beta}{2}$	$\beta < 1$	0

n_a is the number of elements in cluster a

The Davies-Bouldin Index (DBI) was presented in the reference [30], this index evaluates the correlation between the dispersion in the clusters and the distance between the centers of the clusters. As indices MIA and CDI, a lower value of this index indicates well-separated clusters and compacted clusters.

The Dunn Index (DI) is composed of two metrics, the first is the minimum distance between clusters as a metric of separation and the higher this value is, better separated the clusters are. The second metric is the maximum diameter of all clusters as a metric of compactness. The relation between these two metrics is the DI. As is possible to see, as bigger is the distance between clusters and as lower is the diameter of the clusters the DI increases, so a bigger value of this index is better. The mathematical formulation of this index is presented in [31].

The Calinski-Harabasz Index (CHI) takes into account two metrics, the first is a between-cluster sum of squares (BGSS), reflecting the separation between clusters and the within-cluster sum of squares (WGSS) reflecting the compaction of the clusters. The CHI is calculated by the relation of BGSS and WGSS weighted by the number of elements and clusters. The mathematical formalization of this index can be seen in reference [32].

5. Classification

Another proposal of this study is to use the consumption pattern of the customers, discovered in the clustering phase, to classify new customers in their correct consumption class. With this knowledge is possible to optimize contractual parameters by choosing tariff structures that can improve the customer savings.

The main objective of this phase is the generation of rules that can classify customers by their consumption patterns. For that, the load diagrams need to be represented by simpler forms, such as load shape indices. By representing each load diagram in a simpler form, the classification algorithm will generate simpler rules to classify the new customers. A simpler representation can also help to improve the time processing during the training and prediction phase of the classification algorithm. These indices were proposed and used in several works and it is possible to see some of them in the references [19-20], [33-35]. The indices used in this paper are summarized in Table 2.

To classify the load diagrams, it was used the C4.5 algorithm [36], which creates a decision tree capable to classify new customers in predefined classes. Consider the dataset used to train the model as X=[x_1,x_2,… x_n], and x_i consists in a n- dimensional vector with the load shape indices and the class of each load diagram obtained in the clustering phase x_i=[f_1,f_2,f_3,c]. The algorithm uses the vector X to create a decision tree, based on the attributes that can provide the highest information gain to split the dataset. The decision tree can be resumed in a set of rules and based on these, it is possible to classify new customers just with their historical data, avoiding reprocessing the database and performing a new clustering phase.

Table 2. Load shape indices

Parameter	Definition	Period of definition
Load Factor	$f_1 = \frac{P_{av,day}}{P_{max,day}}$	1 day
Night Impact	$f_3 = \frac{1}{3}\frac{P_{av,night}}{P_{av,day}}$	1 day (8 hours night, form 11 p.m. to 6 a.m.)
Lunch Impact	$f_5 = \frac{1}{8}\frac{P_{av,lunch}}{P_{av,day}}$	1 day (3 hours from 12:00 to 15:00)

6. Case study

The case study was done using data from 194 low voltage (LV) customers. Firstly, the dataset was treated to remove inconsistencies, for then apply clustering algorithms to extract load patterns. After the clustering phase, the classification algorithm was applied to identify each customer class in the dataset and obtain information for classifying new customers.

6.1. Pre-processing

The real dataset was analysed aiming to identify missing values. Missing values in databases can occur due to several causes, among them, the most common is the communication failure between the measuring equipment and the database. The database was verified with the intent to remove these inconsistencies to avoid wrong interpretation in the mining phase. It was identified seven customers with energy consumption equal to zero for all the acquisition period. These customers were removed from the analysis to avoid a wrong interpretation of the dataset and to avoid the detection of nonexistent patterns. Short communication faults (missing values up to one hour) were corrected by using the average of neighbour points to maintain the consistency of the load diagrams. Figure 3 shows an example of missing value correction, where the blue continuous line represents the original load data and the red dotted line is the corrected data. These two data treatments have improved the dataset avoiding information loss.

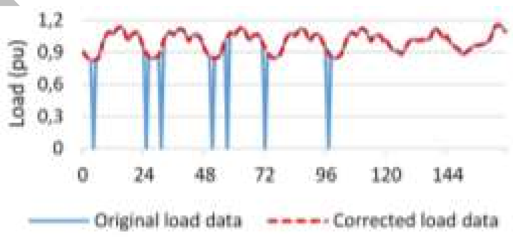

Fig. 3. Lack of values correction.

Analyzing the dataset, it was noted that there were differences between energy consumption amounts and patterns during the weekdays, weekends, and holidays. In Figure 3 it is possible to see (from hour 120) the difference between consumption patterns of weekends and weekdays, where hour 0 represents the first hour of Monday. Due to this difference, it was necessary to separate the data into three categories named by working days, Saturdays, and Sundays/holidays. The third category was composed of Sundays and holidays because the consumption pattern was very similar for those days. With the previous analysis, it was concluded that the consumption pattern for LV customers is highly influenced by economic activities. For example, the consumption pattern for a house on Mondays (where residents are working during the day and are at home at night) tends to be different from that on Saturday (where the majority of the customers are at home during the day and night).

By separating the data into three categories, a data reduction phase was applied to the dataset to reduce the amount of load diagrams and obtain the representative data. The representative load diagrams were obtained by the average of the daily load diagrams of each customer for each category (working days, Saturdays, and Sundays/holidays). This phase is necessary to analyze just relevant data and improve the time in clustering analysis.

After obtained three load diagrams for each customer, a normalization was applied to compare the consumption pattern and not the consumption value. The normalization applied to the dataset was the min-max normalization. This normalization changes the scale of the data to a [0,1] range. It also helps improve the clustering analysis and the interpretation of the load diagrams.

The initial dataset, as all real datasets, was polluted by noising data, so it was cleaned. After that, the data was separated into three categories to improve its interpretation. By separating the dataset in three sub-datasets, a data reduction was applied in an attempt to obtain relevant data for the clustering analysis. The data was normalized to compare the consumption pattern and not the consumption amount of each customer. At the end of the process, the three sub-datasets were passed to a common format composed of 194 LV customers

6.2. Clustering

With the dataset treated in the pre-processing step, the possibility to discover wrong patterns was reduced, so the clustering analysis was applied to identify the Typical Load Profile (TLP). To improve the pattern extraction and avoid wrong interpretation, seven clustering algorithms were applie: K-means, K-medoids, G-means, Average link, Single link, Complete link, and Ward link.

To test the effectiveness of the algorithms, each one was applied to the dataset by dividing the dataset from two clusters to thirteen clusters. Each cluster was evaluated by five clustering indices (MIA, CDI, DBI, DI, and CHI) in order to identify the best number of clusters for each clustering algorithm as well as to find out the best partition.

Each cluster result was evaluated visually to avoid the discovery of clusters with outliers or clusters with almost the same TLP. It was noted that all algorithms found clusters with outliers or clusters with similar TLPs when the dataset was divided into five or more clusters.

Figure 4 a) shows the result of the algorithms when it was used the MIA index. For this index, the algorithm SL presented the best result with thirteen clusters. However, analysing the partitions, it was verified that the algorithm has divided the dataset poorly and did not provide good TLPs, indeed. The AL algorithm shows the second-best result but has the same problem of the SL, by presenting poorly partitions. For the dataset tested in this analysis, the hierarchical algorithms do not found good partitions of the dataset, just the WL algorithm showed reasonable result.

It is possible to see in Figure 4 b) that K-means algorithm provides good dataset partitions, for CDI index, with just two clusters, the algorithms K-medoids and the hierarchical Ward link have shown good results too. The algorithm AL showed better results with three partitions and the SL algorithm provided the best result for the rest of the number of the partitions

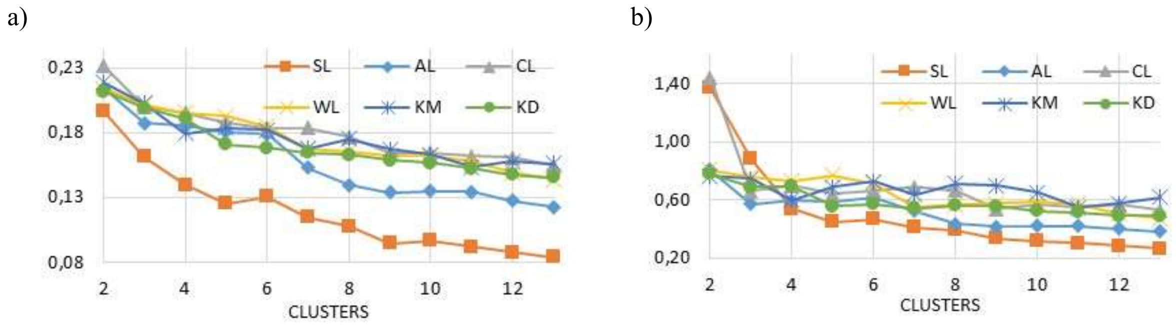

Fig. 4. a) MIA Index results; b) CDI Index results

The K-means algorithm has obtained the best result with two partitions for the DBI (Figure 5 a)), following by the Single Link algorithm with ten partitions. For the Dunn Index, Figure 5 b), the K-means algorithm showed the best number of partitions with two clusters

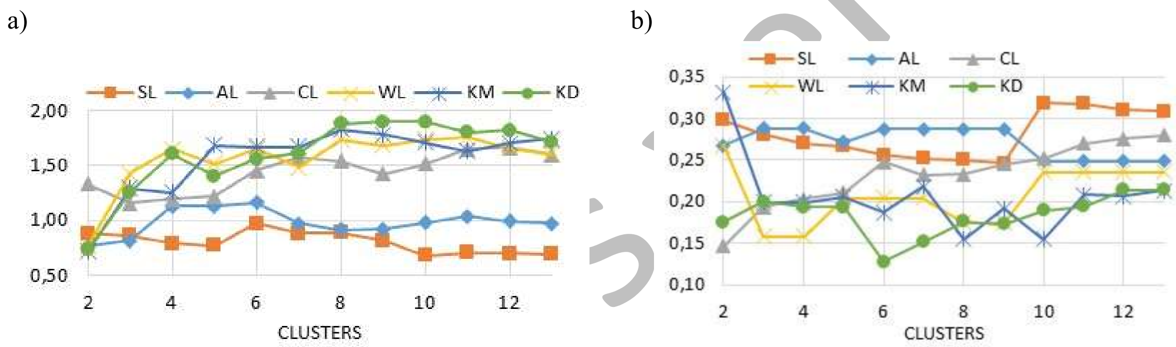

Fig. 5. a) DBI Index results; b) Dunn Index results

When using Calinski-Harabasz index, Figure 6, the K-means provided the best data partitions with three clusters, followed by the partitional algorithms K-medoids with also three partitions.

The G-means algorithm is an automatic clustering algorithm, it means that the algorithm will find the best partitions automatically. The G-means algorithm has divided the dataset into thirteen clusters and did not give great information about the load patterns, once some of the clusters were similar to each other.

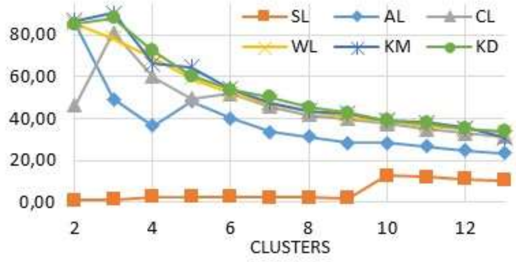

Fig. 6. Calinski-Harabasz index results

The algorithm that provided the best solution was, in fact, the partitional algorithm K-means (using DBI, DI, and CDI evaluation measures indices), followed by Single Link, Average Link, and K-medoids. Unfortunately, the hierarchical clustering algorithms did not find good partitions and it was not possible to consider their partitions in the analysis. Just the hierarchical algorithm, Ward Link, showed good partitions of the dataset.

In Figure 7 it is possible to see the result of the K-means algorithm with two partitions. It is possible to identify one pattern in cluster 1. On the other hand, it is difficult to identify the pattern in cluster 2 (Figure 7 b)), so it was necessary to divide the dataset in three clusters.

In Figure 8 it is possible to observe three distinct TLPs and Figure 9 depicts the result for four partitions of the dataset, it can be seen four well separated TLPs.

Considering the indices results and the analysis of the partitions, the best number of partitions is three, using the K-means algorithm. To proceed with the analysis and the classification phase is necessary to identify the TLP of each cluster. The TLPs were obtained by the average of the load diagrams in each cluster. The TLPs for working days, Saturdays, and Sundays/holidays can be seen in Figure 10 a), b) and c), respectively. It is possible to identify three distinct consumption patterns (3 clusters).

In Figure 10 a) it is possible to see an anomalous pattern with most of the consumption during the period of the night (cluster 1) and a residual consumption during the day. The second pattern obtained, cluster 2, it is considered a residential pattern and it is possible to see a residual energy consumption during the early hours of the day. This consumption increases in the morning and maintains constantly during the day and at night is possible to see a peak consumption. The third cluster was considered as a commercial activity since then presents a residual consumption in the early hours of the day. In working hours, the consumption increases and it is constant during the day and then, at the end of the day, the consumption starts to decrease.

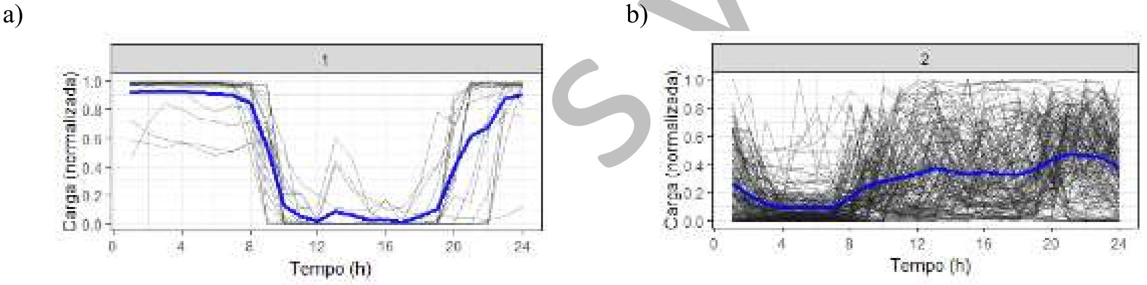

Fig. 7. a) and b) K-means clustering results for 2 clusters

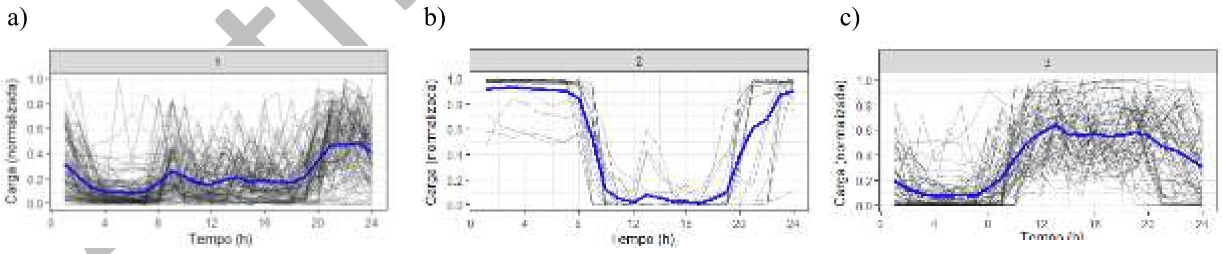

Fig. 8. a), b) and c) K-means clustering results for 3 clusters

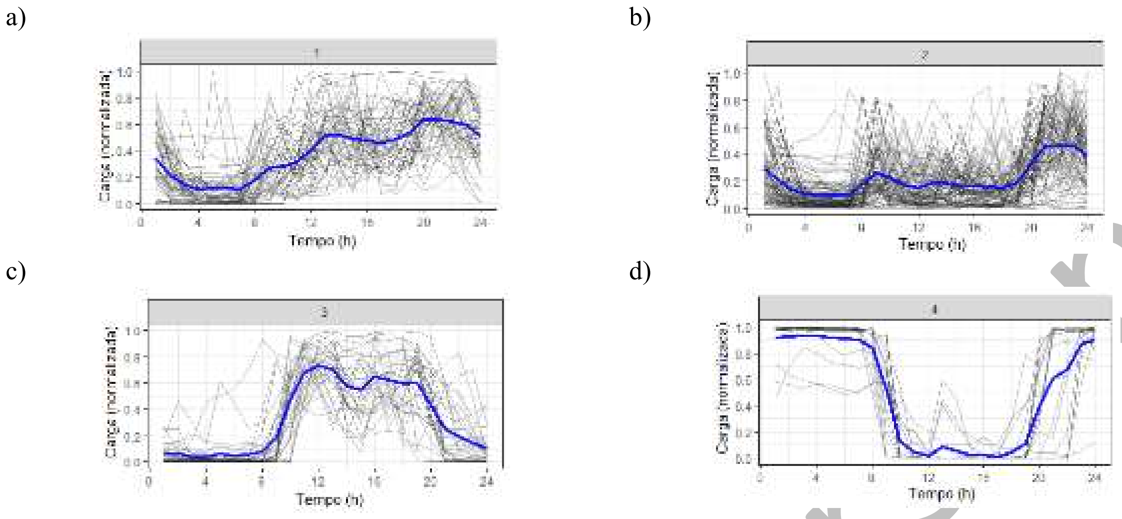

Fig. 9. a), b), c) and d) K-means clustering results for 4 clusters

Comparing now the consumption patterns during the working days (Figure 10 a)), Saturdays (Figure 10b)), and Sundays/holidays (Figure 10 c)) it is possible to verify the difference among patterns, mainly on cluster 2, in the residential consumption profile. It is possible to note the high influence of economic activities, and their impact on consumption habits. Cluster 1 was not influenced by the economic activities and on the opposite, cluster 3 was influenced by them and the profile change to the working days, Saturdays, and Sundays.

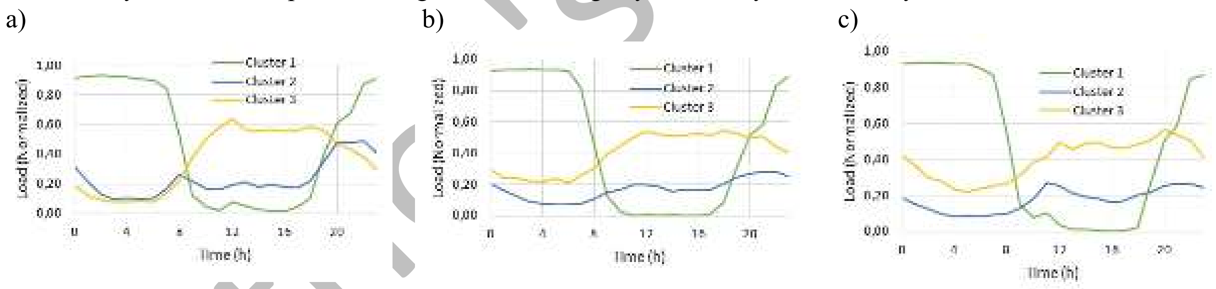

Fig. 10. a) Typical Load Profiles for working days; b) Typical Load Profiles for Saturdays; c) Typical Load Profiles for Sundays

6.3. Classification

To finish the DM cycle and to use the knowledge extracted from the dataset, in the clustering phase, a classification model was implemented to classify electricity customers based on their electricity consumption data. The clustering phase is considered unsupervised learning, where the algorithms "learn" based on the data and extract the relevant information. In the classification step, the knowledge obtained with clustering algorithms is used to train the classification model and predict the classes for new customers.

The classification framework used in this analysis can be observed in Figure 11, and it is divided into four steps. First, the representative daily diagrams are obtained in clustering analysis and each customer class. Each load diagram of each customer is translated into a set of load shape indices, in an attempt to provide a simpler representation of each customer. The load shape indices used in this paper are presented in Table 2.

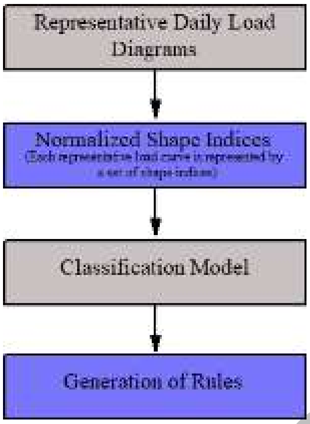
Representative Daily Load Diagrams
Normalized Shape Indices
(Each representative load curve is represented by a set of shape indices)
Classification Model
Generation of Rules

Fig. 11. Costumers' classification framework

To classify the load diagrams, it was used the C4.5 algorithm. The algorithm creates a decision tree to classify new customers based on a set of rules. To train the algorithm were selected randomly 2/3 of the input data. At the end of the training phase the algorithm will generate the rules, to classify a new consumer. The algorithm has generated ten rules, based on three load shape indices, to classify the new customers. The rules can be seen in Table 3.

To test the classification rules provided by the algorithm, 1/3 of the costumers in the dataset were selected randomly and their classes were predicted according to the defined rules. The confusion matrix of the classification can be seen in Table 4, where 64 customers were tested and the algorithm predicted 60 customers to their correct classes. One customer was classified in class 1 (or cluster 1) and the real class was in fact class 2. Another one was classified in class 3, belonging to class 2. Two customers were classified as class 2 but they belong to class 3.

Table 3. Obtained classification rules set for working days, from C.5 algorithm

Rules	Cluster
If $f_3 > 0{,}444$ & $f_1 \leq 0{,}363$	**2**
If $f_3 > 0{,}444$ & $f_1 > 0{,}363$	**1**
If $f_3 \leq 0{,}444$ & $f_5 \leq 0{,}084$	**2**
If $f_3 \leq 0{,}444$ & $f_5 > 0{,}084$ & $f_3 \leq 0{,}114$	**3**
If $f_3 \leq 0{,}444$ & $f_5 > 0{,}084$ & $f_3 > 0{,}114$ & $f_1 \leq 0{,}328$	**2**
If $f_3 \leq 0{,}444$ & $f_5 > 0{,}084$ & $f_3 > 0{,}114$ & $f_1 > 0{,}328$ & $f_3 > 0{,}280$ & $f_1 \leq 0{,}480$	**2**
If $f_3 \leq 0{,}444$ & $f_5 > 0{,}084$ & $f_3 > 0{,}114$ & $f_1 > 0{,}328$ & $f_3 > 0{,}280$ & $f_1 > 0{,}480$	**3**
If $f_3 \leq 0{,}444$ & $f_5 > 0{,}084$ & $f_3 > 0{,}114$ & $f_1 > 0{,}328$ & $f_3 \leq 0{,}280$ & $f_1 > 0{,}404$	**3**
If $f_3 \leq 0{,}444$ & $f_5 > 0{,}084$ & $f_3 > 0{,}114$ & $f_1 > 0{,}328$ & $f_3 \leq 0{,}280$ & $f_1 \leq 0{,}404$ & $f_3 \leq 0{,}213$	**3**
If $f_3 \leq 0{,}444$ & $f_5 > 0{,}084$ & $f_3 > 0{,}114$ & $f_1 > 0{,}328$ & $f_3 \leq 0{,}280$ & $f_1 \leq 0{,}404$ & $f_3 > 0{,}213$	**2**

Table 4. Confusion matrix

Real Class	**Predicted**		
	1	**2**	**3**

1	**4**	0	0
2	1	**32**	1
3	0	2	**24**

The classification tree, generated by the model, can be seen in Figure 12, where all the classification rules were extracted. It is possible to see the number of customers predicted in each class and the customers wrong predicted.

Overall, the classification model has obtained 93.75% of accuracy for the working days, 85.71% for Saturdays and 87.77% for Sundays/holidays. The model proves to be efficient, once, for all the analysed days the model has obtained good results in predict the customers' classes.

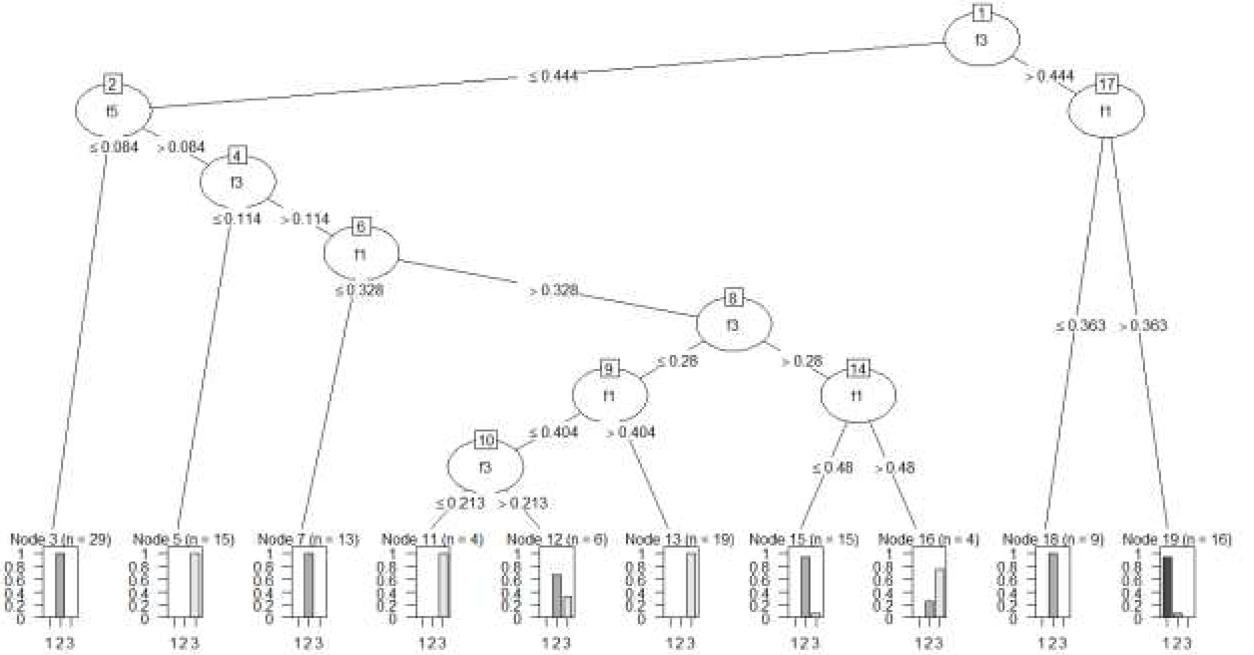

Fig. 12. Example of the obtained decision tree, with the classification rules for the working days.

7. Conclusion

This paper uses a methodology using KDD and DM for the characterization of low voltage customers, based on their historical consumption data. Concepts of the KDD process were addressed and developed in this work, such as data selection, data pre-processing (thorough actions of cleaning, reduction, and transformation), data mining (thorough clustering and classification). Lastly, the model was evaluated.

The performance of 7 clustering algorithms was evaluated through 5 clustering validity indices. The best clustering algorithm were identified and also the best number of clusters to characterize low voltage customers. A classification algorithm was implemented in order to classify new electricity customers, based on their historical consumption data, in one of the obtained clusters.

The model proposed in this paper proved to be an efficient framework in clustering analysis and classification, obtaining acceptable performance. The methodology adopted in this paper can be applied to other datasets to evaluate the performance of the algorithms.

As future work it is proposed improvements in the G-means algorithm, in order to compare its performance against other algorithms. It is also proposed the implementation of another automatic clustering algorithm, the X-means algorithm, and compare the results with G-means and some classical techniques, such as K-means and hierarchical methods.

Acknowledgements

This work has received funding from FEDER Funds through COMPETE program and from184 National Funds through FCT under the project BENEFICE–PTDC/EEI-EEE/29070/2017 and185 UIDB/00760/2020 under CEECIND/02814/2017 grant.

References

1. SRG/AID, "ANEEL atua para implementar a decisão do preço horário no setor," Aug. 01, 2019. [Online]. Available: https://www.aneel.gov.br/sala-de-imprensa-exibicao-2/-/asset_publisher/zXQREz8EVlZ6/content/aneel-atua-para-implementar-a-decisao-do-preco-horario-no-setor/656877?inheritRedirect=false. [Accessed: 02-Feb-2020].

2. M. Ashouri, F. Haghighat, B. C. M. Fung, A. Lazrak, and H. Yoshino, "Development of building energy saving advisory: A data mining approach," Energy Build., vol. 172, pp. 139–151, Aug. 2018.

3. A. M. S. Ferreira, C. A. M. T. Cavalcante, C. H. O. Fontes, and 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., vol. 53, pp. 824–831, Dec. 2013.

4. A. Capozzoli, M. S. Piscitelli, and S. Brandi, "Mining typical load profiles in buildings to support energy management in the smart city context," Energy Procedia, vol. 134, pp. 865–874, Oct. 2017.

5. A. I. Saleh, A. H. Rabie, and K. M. Abo-Al-Ez, "A data mining based load forecasting strategy for smart electrical grids," Adv. Eng. Informatics, vol[6] A. Goia, C. May, and G. Fusai, "Functional clustering and linear regression for peak load forecasting," Int. J. Forecast., vol. 26, no. 4, pp. 700–711, Oct. 2010.

6. A. Goia, C. May, and G. Fusai, "Functional clustering and linear regression for peak load forecasting," Int. J. Forecast., vol. 26, no. 4, pp. 700–711, Oct. 2010.

7. C. H. Jin, G. Pok, I. Paik, and K. H. Ryu, "Short-term electricity load and price forecasting based on clustering and next symbol prediction," IEEJ Trans. Electr. Electron. Eng., vol. 10, no. 2, pp. 175–180, Mar. 2015.

8. P. R. S. Jota, V. R. B. Silva, and F. G. Jota, "Building load management using cluster and statistical analyses," Int. J. Electr. Power Energy Syst., vol. 33, no. 8, pp. 1498–1505, Oct. 2011.

9. T. Teeraratkul, D. O'Neill, and S. Lall, "Shape-Based Approach to Household Electric Load Curve Clustering and Prediction," IEEE Trans. Smart Grid, vol. 9, no. 5, pp. 5196–5206, Sep. 2018.

10. X. Fu, X.-J. Zeng, P. Feng, and X. Cai, "Clustering-based short-term load forecasting for residential electricity under the increasing-block pricing tariffs in China," Energy, vol. 165, pp. 76–89, Dec. 2018.

11. R. Pereira et al., "A fuzzy clustering approach to a demand response model," Int. J. Electr. Power Energy Syst., vol. 81, pp. 184–192, Oct. 2016.

12. R. Li, F. Li, and N. D. Smith, "Multi-Resolution Load Profile Clustering for Smart Metering Data," IEEE Trans. Power Syst., vol. 31, no. 6, pp. 4473–4482, Nov. 2016.

13. G. Chicco, R. Napoli, P. Postolache, M. Scutariu, and C. Toader, "Customer characterization options for improving the tariff offer," IEEE Trans. Power Syst., vol. 18, no. 1, pp. 381–387, Feb. 2003.

14. S. Ramos, Z. Vale, J. Santana, and F. Rodrigues, "Use of data mining techniques to characterize MV consumers and to support the consumer supplier relationship," Proc. 6th WSEAS Int. Conf. Power Syst. Lisbon, Port., pp. 22–24, 2006.

15. F. Biscarri, I. Monedero, A. García, J. I. Guerrero, and C. León, "Electricity clustering framework for automatic classification of customer loads," Expert Syst. Appl., vol. 86, pp. 54–63, Nov. 2017.

16. F. McLoughlin, A. Duffy, and M. Conlon, "A clustering approach to domestic electricity load profile characterisation using smart metering data," Appl. Energy, vol. 141, pp. 190–199, Mar. 2015.

17. G. J. Tsekouras, N. D. Hatziargyriou, and E. N. Dialynas, “Two-Stage Pattern Recognition of Load Curves for Classification of Electricity Customers,” IEEE Trans. Power Syst., vol. 22, no. 3, pp. 1120–1128, Aug. 2007.

18. S. Ramos, J. M. Duarte, F. J. Duarte, and Z. Vale, “A data-mining-based methodology to support MV electricity customers’ characterization,” Energy Build., vol. 91, pp. 16–25, Mar. 2015.

19. V. Figueiredo, F. Rodrigues, Z. Vale, and J. B. Gouveia, “An Electric Energy Consumer Characterization Framework Based on Data Mining Techniques,” IEEE Trans. Power Syst., vol. 20, no. 2, pp. 596–602, May 2005.

20. M. Piao and K. H. Ryu, “Local characterization-based load shape factor definition for electricity customer classification,” IEEJ Trans. Electr. Electron. Eng., vol. 12, pp. S110–S116, Jun. 2017.

21. J. Han, M. Kamber, and J. Pei, Data mining concepts and techniques, 3rd ed. Waltham, 2011.

22. J. A. Hartigan and M. A. Wong, “Algorithm AS 136: A K-Means Clustering Algorithm,” Appl. Stat., vol. 28, no. 1, p. 100, 1979.

23. T. Zhang, G. Zhang, J. Lu, X. Feng, and W. Yang, “A New Index and Classification Approach for Load Pattern Analysis of Large Electricity Customers,” IEEE Trans. Power Syst., vol. 27, no. 1, pp. 153–160, Feb. 2012.

24. I. Dent, T. Craig, U. Aickelin, and T. Rodden, “Variability of Behaviour in Electricity Load Profile Clustering; Who Does Things at the Same Time Each Day?” 2014, pp. 70–84.

25. R. Granell, C. J. Axon, and D. C. H. Wallom, “Impacts of Raw Data Temporal Resolution Using Selected Clustering Methods on Residential Electricity Load Profiles,” IEEE Trans. Power Syst., vol. 30, no. 6, pp. 3217–3224, Nov. 2015.

26. L. Kaufman and P. J. Rousseeuw, “Clustering by means of Medoids, in Statistical Data Analysis Based on the L1 - Norm and Related Methods,” Y. Dodge, Ed. North- Holland, 1987, pp. 405–416.

27. G. Hamerly and C. Elkan, “Learning the K in K-means,” Neural Inf. Process. Syst., 2003.

28. G. N. Lance and W. T. Williams, “A General Theory of Classificatory Sorting Strategies: 1. Hierarchical Systems,” Comput. J., vol. 9, no. 4, pp. 373–380, Feb. 1967.

29. W. H. E. Day and H. Edelsbrunner, “Efficient algorithms for agglomerative hierarchical clustering methods,” J. Classif., vol. 1, no. 1, pp. 7–24, Dec. 1984.

30. D. L. Davies and D. W. Bouldin, “A Cluster Separation Measure,” IEEE Trans. Pattern Anal. Mach. Intell., vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.

31. J. C. Dunn†, “Well-Separated Clusters and Optimal Fuzzy Partitions,” J. Cybern., vol. 4, no. 1, pp. 95–104, Jan. 1974.

32. T. Calinski and J. Harabasz, “A dendrite method for cluster analysis,” Comm. Stat., vol. 3, no. 1, pp. 1–27, 1974.

33. G. Chicco, R. Napoli, F. Piglione, P. Postolache, M. Scutariu, and C. Toader, “Emergent electricity customer classification,” IEE Proc. - Gener. Transm. Distrib., vol. 152, no. 2, p. 164, 2005.

34. S. Ramos and Z. Vale, “Data mining techniques application in power distribution utilities,” in 2008 IEEE/PES Transmission and Distribution Conference and Exposition, 2008, pp. 1–8.

35. M. Bicego, A. Farinelli, E. Grosso, D. Paolini, and S. D. Ramchurn, “On the distinctiveness of the electricity load profile,” Pattern Recognit., vol. 74, pp. 317–325, Feb. 2018.

36. J. R. Quinlan, C4.5: Programs for Machine Learning. San Mateo, Calif: Morgan Kaufmann Publishers Inc., 1993. 30, no. 3, pp. 422–448, Aug. 2016.