DETERMINATION OF ELECTRICITY CONSUMERS' LOAD PROFILES VIA WEIGHTED EVIDENCE ACCUMULATION CLUSTERING USING SUBSAMPLING

Jorge Duarte[1], Ana Fred[2], Fátima Rodrigues[1], João Duarte[1], Sérgio Ramos[1], Zita Vale[1]
[1]GECAD – Knowledge Engineering and Decision Support Group
Instituto Superior de Engenharia do Porto, Instituto Superior Politécnico, Porto, PORTUGAL
{jduarte, fr}@dei.isep.ipp.pt; jmmd@isep.ipp.pt; {sramos, zav}@dee.isep.ipp.pt
www.gecad.isep.ipp.pt/Gecad
[2]Instituto de Telecomunicações, Instituto Superior Técnico, Lisboa, PORTUGAL
afred@lx.it.pt

ABSTRACT

With the electricity market liberalization, the distribution and retail companies are looking for better market strategies based on adequate information upon the consumption patterns of its electricity consumers. A fair insight on the consumers' behavior will permit the definition of specific contract aspects based on the different consumption patterns. In order to form the different consumers' classes, and find a set of representative consumption patterns we use electricity consumption data from a utility client's database and two approaches: Two-step clustering algorithm and the WEACS approach based on evidence accumulation (EAC) for combining partitions in a clustering ensemble. While EAC uses a voting mechanism to produce a co-association matrix based on the pairwise associations obtained from *N* partitions and where each partition has equal weight in the combination process, the WEACS approach uses subsampling and weights differently the partitions. As a complementary step to the WEACS approach, we combine the partitions obtained in the WEACS approach with the ALL clustering ensemble construction method and we use the Ward Link algorithm to obtain the final data partition. The characterization of the obtained consumers' clusters was performed using the C5.0 classification algorithm. Experiment results showed that the WEACS approach leads to better results than many other clustering approaches.

KEY WORDS

Electricity Markets, Load Profiles, Clustering.

1. Introduction

Nowadays, in some European countries, all the consumers are able to buy electric power from the new private suppliers – the retail companies. In the last decade, we have been assisting to the end of electric regulated monopolies and we have definitively entered in the liberalized environment of the electrical market.

Therefore, considering this new context, the knowledge about consumer's consumption patterns (daily load profiles) will be crucial for the accomplishment of agreements on the price of electric power between consumers and suppliers, the definition of marketing policies and innovative contracts and services. In conclusion, the knowledge of the consumer's daily power consumption profile is extremely important to support the relationship between electrical consumers and suppliers.

The definition of consumer classes can be conveniently extracted by knowing the consumer's real electrical behaviour and also by additional external features information, such as weather data, type of activity, contracted power value, consumed energy and tariff type. These consumer's classes can be obtained using clustering approaches. One of the important tools defined using this data is the load profile for different consumer classes. A load profile can be defined as a pattern of electricity demand for a consumer, or group of consumers, over a given period of time.

In this paper we aim to identify the best representative load diagrams of MV electrical consumers, using a given data sample from a monitoring campaign, carried out by the Portuguese utility.

Clustering can be defined as the process of grouping data into distinct classes or clusters based on an appropriate notion of closeness or similarity among data. Even though there are hundred of clustering algorithms in the literature [1-3], no single algorithm can effectively find by itself all types of cluster shapes and structures. With the objective to solve this limitation, some combination clustering ensemble approaches have been proposed [4-10] based on the idea of combining the results of a clustering ensemble into a final data partition.

We build on the work by Fred et al [4,11,12], on evidence accumulation clustering. The idea of evidence accumulation-based clustering is to combine the results of multiple clusterings into a single data partition, by viewing each clustering result as an independent evidence of data organization. EAC takes the co-occurrences of pairs of patterns in the same cluster to combine the results of a cluster ensemble into a single final data partition. The *N* data partitions of *n* patterns are mapped into an $n \times n$

co-association matrix, $Co\_assoc(i,j) = votes_{ij} / N$, where $votes_{ij}$ is the number of times the pattern pair (i,j) is assigned to the same cluster among the N clusterings. Finally, by applying a clustering algorithm to the co-association matrix we obtain the final combined data partition. Duarte et al. proposed the WEAC approach [13-15], also based on evidence accumulation clustering. WEAC uses a weighted voting mechanism to integrate the partitions of the clustering ensemble, leading to a weighted co-association matrix (*w_co_assoc* matrix). Two different methods are used to weight each clustering to be incorporated in the *w_co_assoc* matrix.

Duarte et al. tested how subsampling techniques influence the combination results using the WEAC approach (WEAC with subsampling, WEACS) [16]. Partitions in the ensemble were generated by clustering subsamples of the data set. Each subsample has 80% of the elements of the data set. As with the WEAC approach, two different methods are used to weight data partitions in the co-association matrix (*w_co_assoc* matrix): Single Weighted EAC with subsampling (SWEACS) and Joint Weighted EAC with subsampling (JWEACS).

The WEACS approach was evaluated experimentally [16] on synthetic and real data sets, in comparison with the single application of Single Link, Complete Link, Average Link, K-means and Clarans algorithms, with the subsampling version of EAC, and with the graph-based combination methods by Strehl and Gosh (HGPA, MCLA and CSPA). In [16] we show that the WEACS approach obtains for all these data sets better results than all of the other clustering approaches, with an improvement percentage superior to 10%, allowing concluding that this approach is robust and can be followed to obtain good clusterings.

Section 2 summarizes the cluster validity indices used in WEACS. Section 3 summarizes the Two-step algorithm. Section 4 presents the Weighted Evidence Accumulation Clustering with subsampling (WEACS) and the experimental setup used. Section 5 presents the representative load profiles obtained by the application of WEACS approach to an electricity consumption data set and the characterization of the obtained clusters. Finally, section 6 presents the conclusions and some ideas for future work.

2. Cluster Validity Indices

How many clusters are present in the data and how good is the clustering itself are two important questions that have to be addressed in any clustering. Cluster validity indices provide the formal mechanisms to give an answer to these questions. For a summary of cluster validity measures and comparatives studies see for instance [17,18] and the references therein.

There are three approaches to assess cluster validity [19]: external, internal and relative validity indices.

In this paper we make use of a set of internal and relative cluster validity indices, extensively used and referenced in the literature, to assess the quality of data partitions to be included and weighted in the w_co_assoc matrix; external validity criteria is excluded, since it requires the use of a priori information about cluster structure. We used two internal indices, the Hubert Statistic and Normalized Hubert Statistic (NormHub) [20], and fourteen relative indices: Dunn index [21], Davies-Bouldin index (DB) [22], Root-mean-square standard error (RMSSTD) [23], R-squared index (RS) [23], the SD validity index [18], the S_Dbw validity index [18], Caliski & Harabasz cluster validity index [24], Silhouette statistic (S) [25], index I [26], XB cluster validity index [27], Squared Error index (SE), Krzanowski & Lai (KL) cluster validity index [28], Hartigan cluster validity index (H) [29] and the Point Symmetry index (PS) [30].

3. Two-Step Algorithm

The Two-step clustering method is a scalable cluster analysis algorithm designed to handle very large data sets and it can handle continuous and categorical variables or attributes. It requires only one data pass and has two steps: 1) pre-cluster the cases (or records) into many small sub-clusters; 2) cluster the sub-clusters resulting from pre-cluster step into the desired number of clusters. It can also automatically select the number of clusters. More details about this clustering method can be found in [33].

4. Weighted Evidence Accumulation Clustering using Subsampling (WEACS)

The WEACS approach [16] is an extension of the WEAC approach [13-15] by using subsampling in the construction of the cluster ensemble. Subsampling is used in WEACS to produce diversity in the cluster ensemble and to test the robustness of the approach. In fact, other works have shown that the use of subsampling increase diversity in the cluster ensemble leading to more robust solutions [6,8,10]. Both methods extend the EAC technique by weighting differently each data partition in the combination process, based on the quality of these data partitions, as assessed by cluster validity indices. WEACS proposes the assessment of the quality of each data partition by one or more cluster validity indices, determining its weight in the combination process. The aim of this differentiation in the weighting of the data partitions is to fight with what can happen in a simple voting mechanism when a set of poor clusterings can overshadow another isolated good clustering. By weighting the data partitions in the weighted co-association matrix according to the assessment made by cluster validity indices and by assigning higher importance to better data partitions in the clustering ensemble, we expect to obtain better combination results.

Considering n the number of patterns in a data set and given a clustering ensemble $\mathcal{P}=\{P^1, P^2, ..., P^N\}$ with N partitions of n*0.8 patterns produced by clustering

subsamples of the data set, and a corresponding set of normalized indices with values in the interval [0,1] measuring the quality of each of these partitions, the clustering ensemble is mapped into a weighted co-association matrix:

$$w\_co\_assoc(i,j)= \sum_{L=1}^{N} \frac{vote_{Lij}.VI^L}{S(i,j)},$$

where N is the number of clusterings, $vote_{Lij}$ is a binary value, 1 or 0, depending if the object pair (i,j) has co-occurred in the same cluster (or not) in the L^{th} partition, VI^L is the normalized cluster validity index value for the L^{th} partition and $S(i,j)$ is a matrix such that (i,j)-th entry is equal to the number of data partitions from the total N data partitions where both patterns i and j are simultaneous present. The final combined data partition is obtained by applying a clustering algorithm to the weighted co-association matrix. The proposed WEACS approach is schematically described in table 1.

Table 1. WEACS approach

Input:
n – number of data patterns of the data set
$P = \{P^1, P^2, ..., P^N\}$- Clustering Ensemble with N data partitions of n*0.8 patterns produced by clustering subsamples of the data set
$VI = \{VI^1, VI^2, ..., VI^N\}$ - Normalized Cluster Validity Index values of the corresponding data partitions
Output: Final combined data partitioning.
Initialization: set *w_co_assoc* to a null $n \times n$ matrix.
1. For L=1 to N
   Update the *w_co_assoc*: for each pattern pair (i,j) in the same cluster, set

$$w\_co\_assoc(i,j)=w\_co\_assoc(i,j)+ \frac{vote_{Lij}.VI^L}{S(i,j)}$$

   $vote_{Lij}$ - binary value (1 or 0), depending if the object pair (i,j) has co-occurred in the same cluster (or not) in the L^{th} partition
   VI^L - the normalized cluster validity index value for P^L
   $S(i,j)$ - number of data partitions where patterns i and j are present
2. Apply a clustering algorithm to the *w_co_assoc* matrix to obtain the final data partition

In WEACS we used two different approaches of weighting each data partition:

1. Single Weighted EAC with Subsampling (SWEACS), where the quality of each data partition is assessed by a single normalized relative or internal cluster validity index, and each vote in the *w_co_assoc* matrix is weighted by the value of this index: $VI^L = norm\_validity(P^L)$
2. Joint Weighted EAC with Subsampling (JWEACS), where the quality of each data partition is assessed by a set of relative and internal cluster validity indices, and each vote in the *w_co_assoc* matrix being weighted by the overall contributions of these indices:

$$VI^L = \sum_{ind=1}^{NInd} \frac{norm\_validity_{ind}(P^L)}{NInd}$$

where $NInd$ is the number of cluster validity indices used, and $norm\_validity_{ind}(P^L)$ is the value of the ind^{th} validity index over the partition P^L.

In our experiments, we used sixteen cluster validity indices that can be seen in the papers referred in section 2.

In the WEACS approach we can use different clustering ensembles construction methods, different clustering methods to obtain the final combined data partition, and, particularly in the SWEACS version, we can use even different cluster validity indices to weight the data partitions. These constitute variations of the approach, taking each of the possible modifications as a configuration parameter of the method. Experimental results in [16] show that although the WEACS leads in general to good results, no individual configuration tested led consistently to better best results in all data sets as compared to the subsampling versions of EAC, HGPA, MCLA and CSPA methods.

To solve this problem we use a complementary step to the WEACS approach. It consists in combining the partitions obtained in the WEACS approach with the ALL clustering ensemble construction method. These data partitions are combined using the EAC approach and the final data partition (P*) is obtained by applying the Ward Link algorithm to this new co-association matrix.

4.1 Experimental Setup

4.1.1 Construction of Clustering Ensembles

There are many ways to produce clustering ensembles. In our experiments we produced clustering ensembles using a single algorithm (Single-Link (SL), Complete-Link (CL), Average-Link (AL), K-means and Clarans (CLR)) with different parameters values and/or initializations, and using multiple clustering algorithms with multiple parameters values and/or initializations. Particularly, each clustering algorithm makes use of different values of k and K-means and Clarans in addition make use of different initializations of clusters centers. We explore also a clustering ensemble that includes all the partitions produced by all the clusterings algorithms (ALL).

4.1.2 Normalization of Cluster Validity Indices

Some indices are intrinsically normalized but others are not. In this work we use two indices intrinsically normalized and fourteen that are not. The Normalized Hubert Statistic and Silhouette index are normalized between [-1,1] but we only consider values between [0,1]. We use two internal validity indices and fourteen relative validity indices. For some indices the best result is the highest value and for others the lowest value. For indices of the first type, when the index only have values greater

than zero, the normalization is made by dividing the value obtained for the index by the maximum value obtained over all partitions (*index_value* = *value_obtained*/*Maximum_value*). For indices of the second type, when the index only have values greater than zero, the normalization is made by dividing the minimum value obtained over all partitions by the partition value obtained for the index. (*index_value* = *Minimum_value*/*value_obtained*). Some other indices increase (or decrease) as the number of clusters increase and it is impossible to find neither the maximum nor the minimum. With this kind of indices, we search the value of *k* at which a significant local change in the value of the index happens. This change appears as a "knee" in the plot and corresponds to the number of clusters underlying the data set. In general, the best value of this kind of indices is not the highest (or lowest) value obtained. Hence, this kind of indices can't be integrated directly in the *w_co_assoc* matrix. The best value of these indices is where the "knee" appears. The value 1 is given to the partition correspondent to the "knee" in the index. To integrate these indices in the co-association matrix we implemented the following approach: run the clustering algorithms varying the number of clusters to be obtained between $[1, k_{maximum}]$ where $k_{maximum}$ is the maximum number of clusters we believe to exist in the data set; then, we have to compare the partition correspondent to the "knee" with each of the other partitions generated by this algorithm. We used an external index, the Consistency index (C_i), proposed in [31] to compare these clusterings. We used this approach to Hubert Statistic, RMSSDT index, RS index and Squared Error index. The expected number of clusters in Hartigan cluster validity index is the smallest k>=1 such that H(k)<=10. Since Hartigan index is not calculated for values of k greater than the expected number of clusters (usually achieve negative values) we have to apply to this index the same procedure applied to the indices based on the "knee" to obtain an index value for partitions with k's greater than the expected number of clusters. Table 2 shows the criteria to achieve the best value with each validity index.

Table 2. Criteria to obtain the best value according to each validity index

Index	Criteria	Index	Criteria
Hubert	"Knee"	RMSSDT	"Knee"
NormHub	Max	RS	"Knee"
Dunn	Max	SD	Min
DB	Min	S_Dbw	Min
CH	Max	SE	"Knee"
S	Max	KL	Maximum
I	Max	H	Smallest k: H(k)<=10
XB	Min	PS	Minimum

4.1.3 Extraction of the Final Combined Data Partition

The obtained co-association matrix (*w_co_assoc)* represents a new similarity matrix between patterns and then we can apply a clustering algorithm to it to achieve the final combined data partition *P**. In our experiments, we assumed that the final number of clusters is known and we used the k-means, SL, AL and Ward's link (WR) algorithms to achieve the final partition. To assess the performance of the combination methods, we compare the final data partitions with ground truth information using the Consistency index (Ci) to compare these partitions.

5. Experimental Results

5.1 Data Selection

Our case study is based on a set of 229 MV consumers from a Portuguese utility. Information on the consumer consumption has been gathered by measurement campaigns carried out by EDP Distribuição – a Portuguese Distribution Company, in the nineties, and this data was used for the purpose of a study demonstration.

The monitoring campaigns were based on a load research project for which a sample population, type of consumers (MV, LV), points of meters installation, sampling cadence (15, 30 … minutes) and total duration (months, years...) of data collection were defined.

The instant power consumption for each MV consumer was collected with a cadence of 15 minutes, by real time meters, which gives 96 values a day for each client, for each day of measurement. The measurement campaigns were made during a period of 3 months in the summer and another 3 months in the winter. For this sample, there is also other kind of information, such as the commercial data related to the monthly energy consumption, the activity code and the contracted power.

In tables 3 and 4, it is possible to analyze the distribution of the sample population according to the contracted power and the activity code.

Table 3. Description of the consumer data set – Contracted Power

Contracted Power (kW)	until 250	251 to 500	501 to 1000	1001 to 1500	Then 1500
Consumers Distribution (%)	52,4	18,3	13,5	7,7	8,1

Table 4. Description of the consumer data set – Activity Code

Activity Code	A	B	C	D	E	F	G	H	I	J	K	L	M	N
Consumers Distribution (%)	2,4	6,3	1,9	0,5	9,6	4,8	4,8	0,5	1	0,5	3,8	1	0,5	1,9
Activity Code	**O**	**P**	**Q**	**R**	**S**	**T**	**U**	**V**	**W**	**X**	**Z**	**AA**	**AB**	**AC**
Consumers Distribution (%)	0,5	0,5	4,8	1	1	12	4,3	1	0,5	1	2,4	21,5	4,3	5,7

5.2 Data Pre-processing

As data is always problematic to handle, a previous data-cleaning phase to detect and correct bad data is indispensable to any Data Mining (DM) process. Starting from the initial databases, we have detected some damaged files and some consumers without registered

values. So, twenty-one consumer's files were removed from the initial sample, remaining 208 consumers to be analyzed. In this data-cleaning phase, we filled missing values of measures using a neural net [32]. These failures can be due to transmission interruptions or damage in the measurement equipment. To estimate missing values we have used a multi layer perceptron (MLP) artificial neural net and historical data of electricity consumption. The neural net was trained starting from the report of each consumer's consumption. In figure 1, we can see an example of consumption estimation. This consumption has two points of measure missing values. By completing this missing data, the errors of the metered load curves are attenuated without making significant alterations in the real measures. After the data completion, we have prepared it for clustering.

Each consumer is represented by his representative daily load curve resulting from elaborating the data from the measurement campaign. For each consumer, the representative load diagram has been built by averaging the load diagrams related to each consumer [34]. A different representative load diagram is created to each one of the loading conditions defined: working days and weekends. Each consumer is defined by a representative daily load curve for each of the loading conditions to be studied separately.

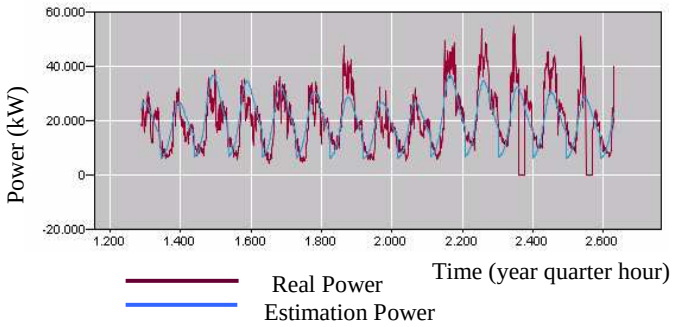

Figure 1- Estimation of a MV consumer consumption using a neural net.

The representative daily load diagram of the m^{th} consumer is the vector $l^{(m)}$:

$$l^{(m)} = \left[l_1^{(m)}, \ldots l_h^{(m)} \right], \; m \in \{1 \ldots M\}, h \in \{1 \ldots H\} \qquad (1)$$

where (m) represents the consumer number in analysis, M represents the number of consumers of the sample and H=96 represents the 15 minute-intervals in a day.

The diagrams were computed using the field-measurements values; therefore they need to be brought together to a similar scale so that pattern may be compared. This is achieved through normalization.

For each consumer the vector represented in (1) was normalized to the [0-1] range by using the peak power of its representative load diagram [32,34]. We choose this kind of normalization to permit the maintenance of curve shape in order to compare the consumption patterns. At this point each consumer is represented by a group H of data consisting of values for 15 minute-intervals, which gives a set of 96 values in the range [0-1].

5.3 Determining of Electricity Consumers' Load Profiles using Two-Step and WEACS approaches

The Two-step and WEACS approaches have been used to group the load patterns on the basis of their distinguishing features. At present, in Portugal, the regulated electrical company has nine consumption patterns. Based on this information we fixed the number of clusters of the final combined data partition in 9 clusters. We obtained the expected 9 clusters for the two different load regimes: work days and weekends.

Figure 2 shows the representative load diagram obtained for each cluster using the Two-step approach and using the measurement power for the working days.

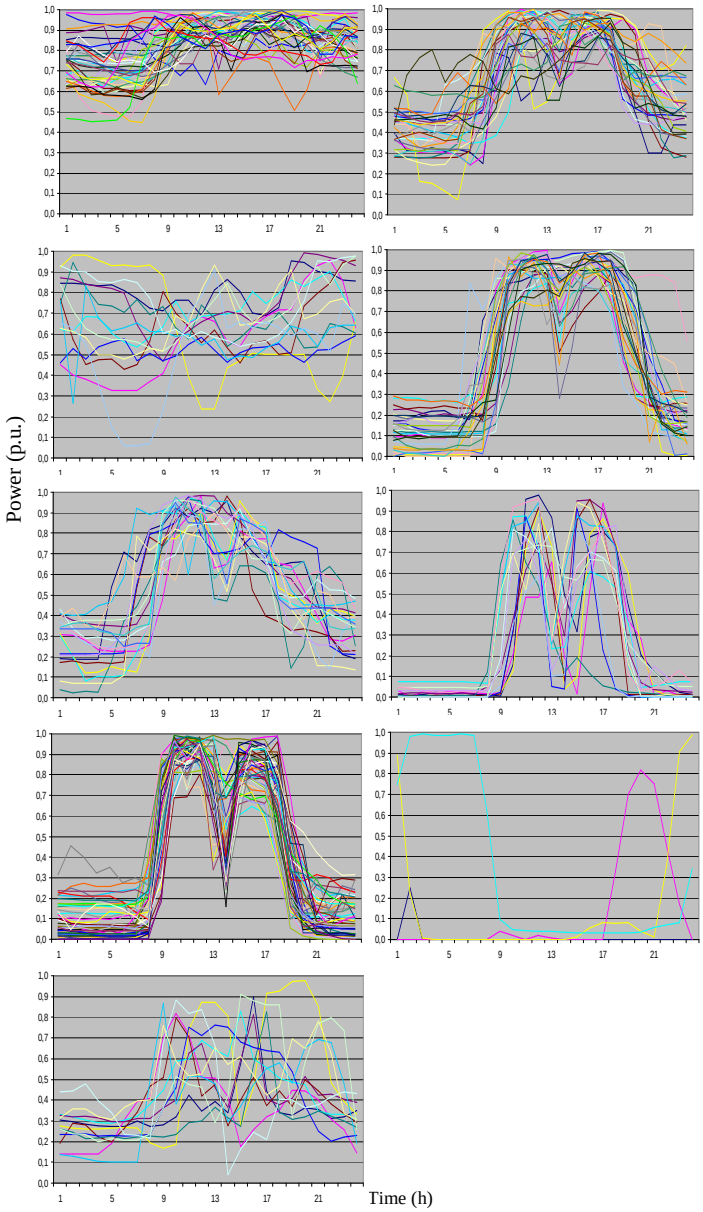

Figure 2- Clusters obtained by Two-step clustering algorithm for working days.

In figure 2, we can see that cluster number 8 contains

four consumers with atypical electric energy consumption. These kinds of atypical consumers (outliers) should be removed from the study so that the characterization results do not become depreciated.

Figure 3 shows the representative load diagram obtained for each cluster using the WEACS approach. Apart from cluster number 9, the WEACS approach separated the consumer population well and the representative load diagrams were created with a distinct load shape. As already mentioned, in this case cluster number 9 should be removed from the study.

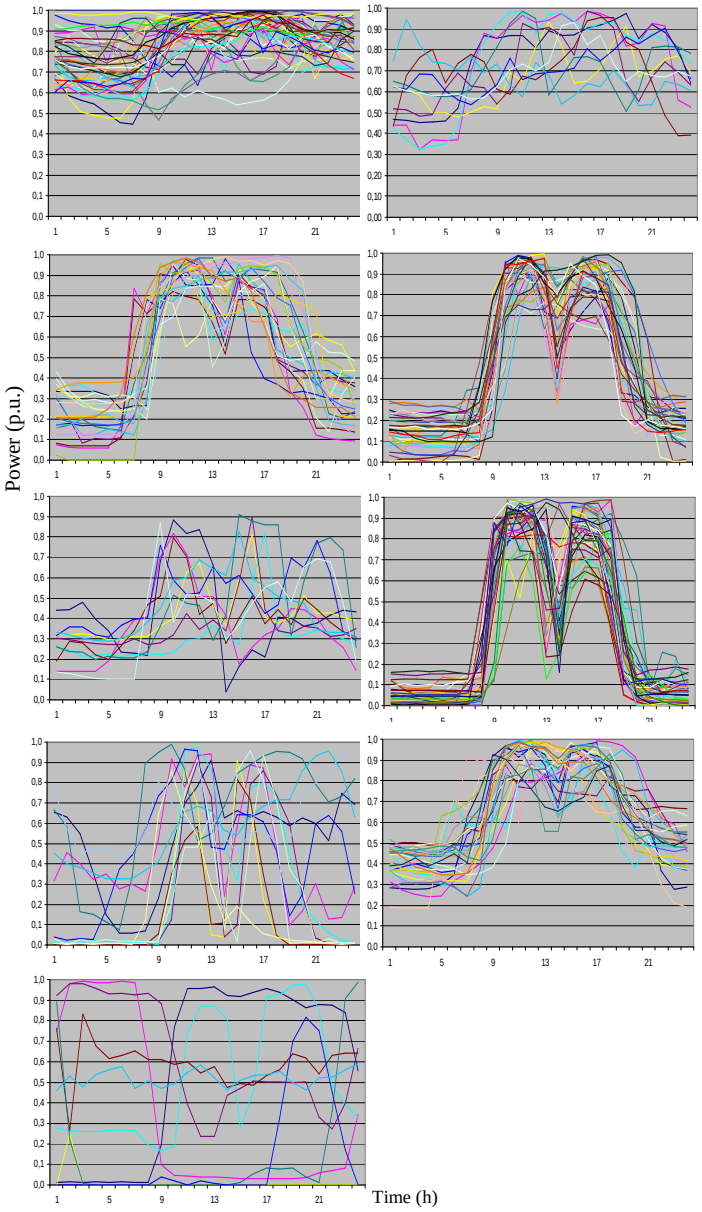

Figure 3- Clusters obtained by WEACS approach for working days.

With the 8 resulting clusters we obtained the representative load diagram for each cluster for working days and weekends by averaging the load diagrams of the clients assigned to the same cluster (figures 4 and 5). Each curve represents the load profile of the corresponding consumer class.

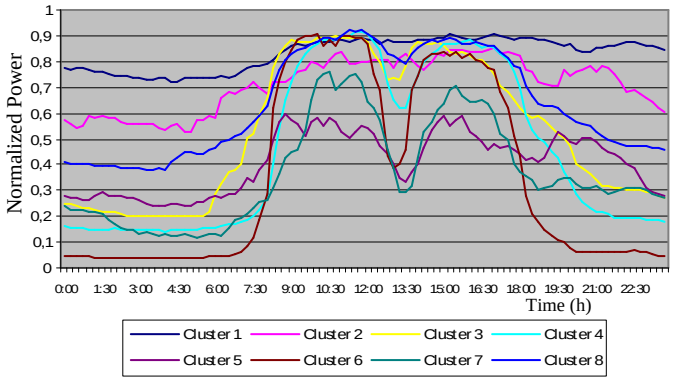

Figure 4- Representative Load Profile for working days clusters

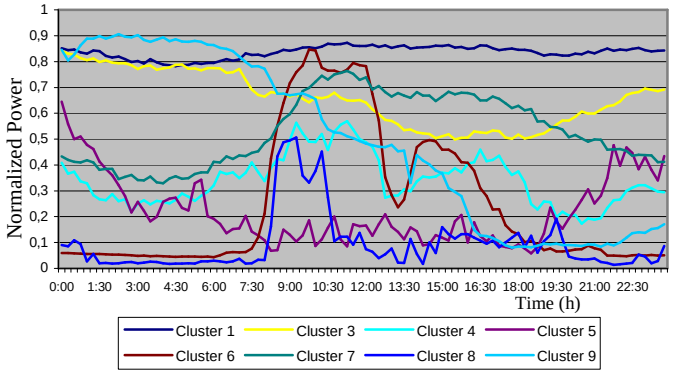

Figure 5- Representative Load Profile for weekend clusters

From the representative load diagrams obtained to each cluster it is possible to see that the WEACS approach has well separated the consumer population, producing representative load diagrams with distinct load shapes. For the characterization of the consumer classes a first trial was made to search for an association between the clusters and the components of the contractual data. Specifically, we searched for an association between the activity type and the hired power of each consumer and the obtained clusters.

From tables 6 (working days clusters) and 7 (weekend clusters) we can see that many of the activity types are present in many clusters, allowing us to conclude that a poor correlation exists between the clusters and the consumers' activity types. Figures 6 (working days clusters) and 7 (weekend clusters) also show that many of the hired powers appear in many clusters. These results show that the contractual data is highly ineffective from the viewpoint of the characterization of the consumers' electrical behaviour

Table 6-Number of consumers of different activity types within each working days clusters.

Activity Type		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Z	AA	AB	AC
Cluster	**1**	-	5	2	1	8	1	1	-	-	-	-	-	-	1	-	-	6	1	-	-	1	-	-	1	-	11	-	1
	2	-	2	-	-	2	-	-	-	1	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	3	-	1
	3	-	-	1	-	3	1	1	-	-	-	-	-	-	1	-	-	-	-	-	6	-	-	-	-	3	6	-	-
	4	1	1	-	-	1	2	-	-	-	1	4	-	-	-	-	-	-	-	-	4	3	-	-	-	2	12	2	1
	5	-	-	-	-	-	-	-	-	-	-	1	-	-	1	-	-	-	-	-	-	-	-	1	-	-	-	-	7
	6	3	2	1	-	4	4	4	1	-	-	-	-	1	-	-	-	2	1	1	9	4	1	-	-	-	5	1	-
	7	-	-	-	-	-	-	1	-	-	-	-	-	-	-	1	1	1	-	1	4	1	-	-	-	-	2	-	1
	8	1	3	-	-	1	2	3	-	-	-	3	-	-	-	-	-	1	-	-	2	-	1	-	-	-	5	4	-
	9	-	-	-	-	1	-	-	-	1	-	-	2	-	-	-	-	-	-	-	-	-	-	-	1	-	1	2	1

Table 7-Number of consumers of different activity types within each weekend clusters.

Activity Type		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Z	AA	AB	AC
Cluster	**1**	1	1	2	1	-	1	2	-	-	-	3	-	-	2	-	-	2	-	-	1	1	1	-	-	1	19	6	1
	2	-	-	-	-	-	-	-	-	-	-	1	-	-	1	-	-	-	-	-	-	2	-	-	1	-	-	1	-
	3	2	6	-	-	5	3	2	-	-	-	2	-	-	1	-	-	3	1	-	2	2	-	-	-	1	9	-	2
	4	1	-	-	-	2	1	3	-	-	1	-	-	-	-	-	-	1	1	-	5	-	-	-	-	-	-	-	-
	5	-	-	-	-	1	-	-	-	-	-	1	-	-	-	1	1	1	-	2	1	-	-	-	-	-	2	1	-
	6	-	1	-	-	3	-	-	1	1	-	-	-	1	-	-	-	-	-	-	5	1	1	-	-	-	1	-	-
	7	-	3	2	-	-	3	1	-	-	-	-	-	-	-	-	-	1	-	-	7	3	-	1	1	3	13	1	8
	8	1	-	-	-	1	-	2	-	-	-	-	-	-	-	-	-	1	-	-	4	-	-	-	-	-	1	-	1
	9	-	2	-	-	8	2	-	-	1	-	1	2	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-

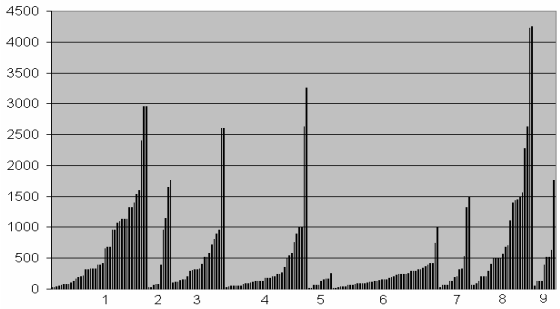

Figure 6- Hired power in working days clusters

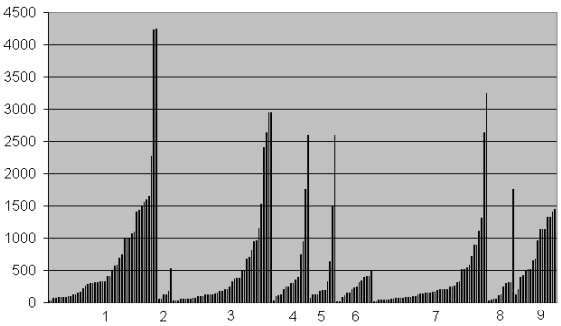

Figure 7- Hired power in weekend clusters

5.4 Classification Model

In order to identify more relevant information on the consumers' consumption behaviour and to describe the consumption patterns of each cluster population we used the C5.0 algorithm to analyze those clusters and to obtain their descriptions based on a set of indices derived from the representative daily load curves.

As the commercial indices had an inexistent correlation with the representative load diagrams of each cluster, we chose other kinds of indices in order to obtain sense rules. As rules must be intelligible, normalized shape indicators were used as attributes in the classification model. Therefore, we extracted several shape indicators that represent the load shape diagram, namely, the load factor (f_1), low factor (f_2), modulation factor (f_3) night impact factor (f_4) and lunch impact factor (f_5), as represented by the vector (f):

$$f = [f_1, f_2, f_3, f_4, f_5] \quad (2)$$

These indices were based on other works related with electrical energy consumption [35].

The classification model uses supervised learning based on the knowledge about the relation between the consumers' load shape indices and the corresponding class obtained through the clustering operation.

Using the representative load diagrams, the load shape indices are computed for each MV consumer. The indices vector (2) that characterizes the representative load diagram shape was used in the classification model as an attribute in order to obtain intelligible rules. These data sets were separated and formed a training set and a test set. The training set used by the classification model has been formed with 2/3 of the data, the remaining 1/3 of the data was used for the test set.

Table 5 presents a rule set example obtained from the

classification algorithm for the working days data set. The obtained rules are simple and easy to understand.

Table 5. Rule set for the working days classification model

If $f_1 \leq 0.57$ and $f_3 \leq 0.2$	then cluster -7
If $f_1 \leq 0.57$ and $f_3 \leq 0.21$ and $f_1 > 0.24$ and $f_4 \leq 0.10$	then cluster -6
If $f_1 \leq 0.57$ and $f_3 \leq 0.21$ and $f_1 > 0.24$ and $f_4 > 0.10$	then cluster -4
If $f_1 \leq 0.57$ and $f_3 \leq 0.21$ and $f_1 > 0.24$ and $f_4 > 0.10$ and $f_1 > 0.44$	then cluster -7
If $f_1 \leq 0.57$ and $f_3 > 0.21$ and $f_5 \leq 0.61$	then cluster -5
If $f_1 \leq 0.57$ and $f_3 > 0.21$ and $f_5 > 0.61$	then cluster -4
If $f_1 \leq 0.57$ and $f_1 > 0.45$ and $f_4 \leq 0.20$	then cluster -4
If $f_1 \leq 0.57$ and $f_1 > 0.45$ and $f_4 \leq 0.20$	then cluster -3
If $f_1 \leq 0.57$ and $f_1 > 0.45$ and $f_4 > 0.20$	then cluster -5
If $f_1 > 0.57$ and $f_1 \leq 0.71$ and $f_4 \leq 0.23$	then cluster -8
If $f_1 > 0.57$ and $f_1 \leq 0.71$ and $f_4 > 0.23$	then cluster -2
If $f_1 > 0.57$ and $f_4 \leq 0.21$	then cluster -2
If $f_1 > 0.57$ and $f_4 > 0.21$	then cluster -1

The classification model considered all the available attributes for each rule, selecting only those that provided larger information gain. This rule set can be used in order to classify new consumers.

6. Conclusions and future work

This paper deals with the clustering of electricity consumers, based on their measured daily load curves.

Two-step cluster algorithm and the WEACS approach were used to obtain the representative load diagrams..

By the observation of the load diagram obtained with each approach, we noticed that the WEACS approach separates the consumer population better than the Two-step cluster algorithm.

The results obtained for both clustering approaches point out that the contractual parameters are poorly connected to the load profiles, so further work was required in order to produce global shape indices able to capture relevant information on the consumers' consuming behaviour.

The characterization of the clusters obtained with WEACS was performed using the C5.0 classification algorithm. Normalized shaped indices were used as attributes in the classification model which generated a rule set easy to understand.

The load profiles will be used to study the best-dedicated tariffs to each consumer class, according to the new rules introduced in the liberalized electricity market.

Following the classification of the consumers into classes, a decision support system will be developed for assisting managers in properly fixing contract details for each consumer classes. This system must be sufficiently flexible to follow the variations in the consumers' load patterns.

Acknowledgements

The authors would like to express their gratitude to EDP Distribuição, the Portuguese Distribution Company, for supplying the data used in this work.
The authors would also like to acknowledge FCT, FEDER, POCTI, POSI, POCI and POSC for their support to R&D Projects and GECAD Unit.

References

[1]A.k. Jain and R.C. Dubes, *Algorithms for Clustering Data* (Prentice Hall, 1988).
[2]A.K. Jain, M.N. Murty, and P.J. Flynn, Data clustering: A review, *ACM Computing Surveys*, 31(3),:264-323, September 1999.
[3]J. Han, M. Kamber, *Data Mining- Concepts and Techniques (*Morgan Kaufmann Publishers, 2001).
[4]A. Fred and A.K. Jain, Combining Multiple Clusterings using Evidence Accumulation. *IEEE Transactions on Pattern analysis and Machine Intelligence*, Vol.27, No.6, June 2005, pp. 835-850.
[5]A. Strehl and J. Ghosh. Cluster ensembles - a knowledge reuse framework for combining multiple partitions, *Journal of Machine Learning Research 3*, 2002.
[6]S.T. Hadjitodorov, L. I. Kuncheva, L. P. Todorova, Moderate Diversity for Better Cluster Ensembles, *Information Fusion*, 2005.
[7]X.Z. Fern, C.E. Broadley, Random projection for high dimensional data clustering: a cluster ensemble approach. *20th International Conference on Machine Learning*, ICML;Washington, DC, 2003, pp. 186-193.
[8]S Monti; P. Tamayo; J. Mesirov; T. Golub, Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data, *Machine learning*, 52, 2003, pp. 91-118.
[9]A. Topchy, B. Minaei-Bidgoli, A.K. Jain, W. Punch, Adaptive Clustering Ensembles. *Proc. Intl. Conf on Pattern Recognition, ICPR'04*, Cambridge, UK, 2004, pp. 272-275.
[10]B. Minaei-Bidgoli, A. Topchy, W. Punch, Ensembles of Partitions via Data Resampling. *Proc. IEEE Intl. Conf. on Information Technology: Coding and Computing, ITCC04*, vol. 2, April 2004, pp. 188-192.
[11]A. Fred and A. K. Jain, Data clustering using evidence accumulation,. *Proc. of the 16th Int'l Conference on Pattern Recognition*, 2002, pp. 276–280.
[12]Fred A., Jain A. K., Evidence accumulation clustering based on the k-means algorithm (*S.S.S.P.R*, T.Caelli et al.,

editor,., Vol. LNCS 2396, Springer-Verlag, 2002, pp. 442–451).

[13] F.Jorge Duarte, Ana L.N. Fred, André Lourenço and M. Fátima C. Rodrigues, Weighting Cluster Ensembles in Evidence Accumulation Clustering. *Workshop on Extraction of Knowledge from Databases and Warehouses, EPIA* 2005.

[14] F.Jorge F.Duarte, Ana L.N. Fred, André Lourenço and M. Fátima C. Rodrigues, Weighted Evidence Accumulation Clustering. *Fourth Australasian Conference on Knowledge Discovery and Data Mining* 2005.

[15]F. Jorge Duarte, Ana L. N. Fred, M. Fátima C. Rodrigues and João Duarte, Evidence Accumulation Clustering using the weight of the cluster ensemble. *International Conference on Knowledge Engineering and Decision Support (ICKEDS-2006)*

[16]F. Jorge Duarte, Ana L. N. Fred, M. Fátima C. Rodrigues and João Duarte, Weighted Evidence Accumulation Clustering using Subsampling. *Sixth International Workshop on Pattern Recognition in Information Systems (PRIS-2006).*

[17]M. Meila and D. Heckerman, "An Experimental Comparison of Several Clustering and Initialization Methods", Proc. 14th Conf. Uncertainty in Artificial Intelligence, p.p. 386-395, 1998.

[18]M. Halkidi, Y. Batistakis, M. Vazirgiannis, *Clustering algorithms and validity measures. Tutorial paper in the proceedings of the SSDBM 2001 Conference*.

[19]Theodorodis, S., Koutroubas, K., *Pattern Recognition (*Academic Press, 1999).

[20]Hubert L.J., Schultz J., Quadratic assignment as a general data-analysis strategy, *British Journal of Mathematical and Statistical Psychology, Vol.29*, 1975, pp. 190-241.

[21]Dunn, J.C., Well separated clusters and optimal fuzzy partitions (J. Cybern, Vol. 4, 1974, pp. 95-104).

[22]Davies, D.L., Bouldin, D.W., A cluster separation measure. *IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 1, No2*, 1979.

[23]S.C. Sharma, *Applied Multivariate Techniques* (John Willwy & Sons, 1996).

[24]Calinski, R.B.& Harabasz, J, A dendrite method for cluster analysis, *Communications in statistics 3*, 1974, pp.1-27.

[25]Kaufman, L. & Roussesseeuw, P., Finding groups in data: an introduction to cluster analysis, *New York, Wiley*, 1990.

[26]U. Maulik and S. Bandyopadhyay, Performance Evaluation of Some Clustering Algorithms and Validity Indices, *IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, no. 12*, 2002, pp. 1650-1654.

[27].Xie, X.L., Beni, G., A Validity Measure for Fuzzy Clustering,, *IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 13*, 1991, pp. 841-847.

[28]W. Krazanowski, Y. Lai, A criterion for determining the number of groups in a dataset using sum of squares clustering, *Biometrics*, 1985, pp. 23-34.

[29]J.A. Hartigan, Statistical theory in clustering, *J. Classification*, 1985, 63-76.

[30]C.H. Chou, M.C. Su, E. Lai, A new cluster validity measure and its application to image compression, *Pattern Analysis and Applications, Vol. 7*, 2004, pp. 205-220.

[31]A. Fred, Finding consistent clusters in data partitions, *Multiple Classifier Systems, Josef Kittler and Fabio Roli editors, vol. LNCS 2096, Springer*, 2001, pp. 309-318.

[32]Sérgio Ramos, Fátima Rodrigues, Raul Pinheiro, Judite Ferreira & Zita Vale, Decision Support System for Improving the Tariff Offer Based on Patterns Extracted from MV Load Diagrams. *Proc. of the International Conference on Knowledge Engineering and Decision Support* (*ICKEDS´06*), pp 107-115, Lisbon, Portugal, May, 2006.

[33]Chiu, T., Fang, D., Chen, J., Wang, Y., and Jeris, C, A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment. *Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 263, 2001.

[34]Jorge Duarte, Fátima Rodrigues, Vera Figueiredo, Zita Vale, M. Cordeiro, Data Mining Techniques Applied to Electric Consumers Characterization, *Proceedings of 7th Artificial Intelligence and Soft Computing*, Canada, Julho de 2003.

[35]Ramos, S., Zita, V., Figueiredo, V., Rodrigues, F. & Pinheiro, R., Characterization of MV Consumers Using Hierarchical Clustering", *Proc. of the International Conference on Knowledge Engineering and Decision Support (ICKEDS´04*), pp 199-204, Porto, Portugal, July, 2004.