Definition of MV Load Diagrams 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

www.gecad.isep.ipp.pt/Gecad
[2]Instituto de Telecomunicações, Instituto Superior Técnico, Lisboa, PORTUGAL

Abstract: - A definition of medium voltage (MV) load diagrams was made, based on the data base knowledge discovery process. Clustering techniques were used as support for the agents of the electric power retail markets to obtain specific knowledge of their customers' consumption habits. Each customer class resulting from the clustering operation is represented by its load diagram. The Two-step clustering algorithm and the WEACS approach based on evidence accumulation (EAC) were applied to an electricity consumption data from a utility client's database in order to form the customer's classes and to find a set of representative consumption patterns. The WEACS approach is a clustering ensemble combination approach that uses subsampling and that weights differently the partitions in the co-association matrix. As a complementary step to the WEACS approach, all the final data partitions produced by the different variations of the method are combined and the Ward Link algorithm is used to obtain the final data partition. Experiment results showed that WEACS approach led to better accuracy than many other clustering approaches. In this paper the WEACS approach separates better the customer's population than Two-step clustering algorithm.

Key-Words: - Electricity markets, load profiles, clustering, weighted evidence accumulation clustering.

1 Introduction

In the regulated power systems, the information about the customer's consumption was important for managing the demand of power, the system planning or definition of better tariffs. Nowadays, with the emergence of competitive electricity markets (deregulated electricity markets) a reduced price paid by customers for electricity consumption is expected. For the retail companies, the knowledge about customer's consumption patterns is very important for the accomplishment of agreements in the price of the electricity between customers and suppliers, the definition of marketing policies and innovative contracts and services. For suppliers who choose a differentiation strategy, the knowledge of the needs of their customers is very important to develop products to suit their preferences. To achieve success in deregulated markets, companies must learn to segment the market and target these segments with the most effective types of marketing methods. One possible method of differentiation is the development of tailored contracts defined according to customer consumption patterns.

One of the important results obtained using this data are the load profiles for different customer classes. A load profile can be defined as a pattern of electricity demand for a customer, or group of customers, over a given period of time. The accurate classification of customer classes and the association of a load profile are essential to support marketing strategies. These customer's classes can be obtained using clustering approaches.

The definition of customer's classes can be extracted by the knowledge of the real customers' electrical behavior and also by additional external features information, such as weather data, activity type, contracted power value, consumed energy and tariff type. The identification of the best representative load diagrams of MV electrical customers is proposed, using a given number of daily load diagrams extracted from a monitoring campaign carried out by the Portuguese utility. For the definition of MV customers' load profiles, the Two-step clustering algorithm and the WEACS approach based on evidence accumulation (EAC) were used.

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 [4-9] based on the idea of combining the results of a clustering ensemble into a final data partition.

The work on evidence accumulation clustering conducted by Fred et al [4-6] has been used as basis for this work. 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 data partitions are mapped into a co-association matrix and the final combined data partition is obtained by applying a clustering algorithm to the co-association matrix.

Duarte et al. proposed the WEAC approach [10-12], 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. In the first method, the Single Weighted EAC (SWEAC), each clustering is evaluated by a relative or internal cluster validity index and the contribution of each clustering is weighted by the value obtained for this index. In the second method, the Joint Weighted EAC (JWEAC), each clustering is evaluated by a set of relative and internal cluster validity indices and the contribution of each clustering is weighted by all results obtained with each of these indices. The final combined partition is obtained by applying a clustering algorithm to the obtained *w_co_assoc* matrix.

Duarte et al. tested how subsampling techniques influence the combination results using the WEAC approach (WEAC with subsampling, WEACS) [13]. 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 [13] 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) [7]. The quality of the final data partition (P^*) is evaluated by calculating the consistency of P^* with ground truth information. It can be seen in [13] that the WEACS approach obtains for almost 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 clustering algorithm. Section 4 presents the Weighted Evidence Accumulation Clustering with subsampling (WEACS) and the experimental setup used. Section 5 presents the obtained representative load profiles obtained by the application of WEACS approach to an electricity consumption data set. 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 [14,15] and the references therein.

There are three approaches to assess cluster validity [16]: external validity indices, where the evaluation of the clustering results is made using a structure that is assumed on the data set (ground truth); internal validity indices, where the assessment of the clustering results is made in terms of the quantities that involve the vectors of the data set themselves; and relative validity indices, where the assessment of a clustering result is made by comparing it to other clustering results, achieved by the same algorithm but with different input parameters.

This work make use of a set of internal and relative clustering 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. Two internal indices, the Hubert Statistic and Normalized Hubert Statistic (NormHub) [17], and fourteen relative indices are used: the SD validity index [15], the S_Dbw validity index [15], Dunn index [18], Davies-Bouldin index (DB) [19], Root-mean-square standard error (RMSSTD) [20], R-squared index (RS) [20], Caliski & Harabasz cluster validity index [21], Silhouette statistic (S) [22], index I [23], XB cluster validity index [24], Squared Error index (SE), Krzanowski & Lai (KL) cluster validity index [25], Hartigan cluster validity index (H) [26] and the Point Symmetry index (PS) [27]. The mathematical equations of all indices used in the WEACS approach are available in [10].

3. Two-Step Algorithm

The Two-step is a scalable cluster algorithm designed to handle very large data sets and that can handle continuous and categorical variables or attributes. It only requires 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 [28]. The Two-step clustering algorithm is included in Clementine version 8.5 [Clementine Data Mining System, web page – http://www.spss.com].

4 Weighted Evidence Accumulation Clustering using Subsampling (WEACS)

The WEACS approach [13] is an extension of the WEAC approach [10-12] 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 [8,9]. 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. While in EAC the N data partitions of n patterns are mapped into an $n\times n$ co-association matrix:

$$Co\_assoc(i,j) = votes_{ij} / N \quad (1)$$

where $votes_{ij}$ is the number of times the pattern pair (i,j) is assigned to the same cluster among the N clusterings, 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 avoid what can happen in a simple voting mechanism when a set of poor clusterings can overshadows another isolated good clustering. Better combination results are expected 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.

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:

$$w\_co\_assoc(i,j)= \sum_{L=1}^{N} \frac{vote_{Lij}.VI^L}{S(i,j)}, \quad (2)$$

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. Information about the normalization of cluster validity indices can be consulted in [13]. 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 algorithm

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 the L^{th} partition
   $S(i,j)$ - 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

WEACS has 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:
$$VI^{L} = norm\_validity\left(P^{L}\right) \qquad (3)$$
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:
$$VI^{L} = \sum_{ind=1}^{NInd} \frac{norm\_validity_{ind}\left(P^{L}\right)}{NInd} \qquad (4)$$
where $NInd$ is the number of cluster validity indices used, and $norm\_validity_{ind}\left(P^{L}\right)$ is the value of the *ind*th validity index over the partition P^L.

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

In the WEACS approach can be used different clustering ensembles construction methods, different clustering methods to obtain the final combined data partition, and, particularly in the SWEACS version, can be used 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 [13] 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 a complementary step is used 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, the clustering ensembles were produced 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. A clustering ensemble that includes all the partitions produced by all the clusterings algorithms (ALL) was also explored.

4.1.2 Extraction of the Final Combined Data Partition

As the obtained co-association matrix (*w_co_assoc)* represents a new similarity matrix between patterns a clustering algorithm can be applied to it to achieve the final combined data partition P^*. In the performed experiments, the final number of clusters is assumed as being known and the k-means, SL, AL and Ward's link (WR) algorithms were used to achieve the final partition. To assess the performance of the combination methods, the final data partitions are compared with ground truth information and the Consistency index (Ci) is used to compare these partitions [6].

5. Experimental Results

5.1 Data Selection

This specific case study is based on a set of 229 MV customers from a Portuguese utility. Information on the customer 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 customers (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 customer 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.

5.2 Data Pre-processing

Starting from the initial databases, some damaged files were detected and some customers without registered values were removed from the initial data sample. So, twenty-one customer’s files were removed from the initial sample, remaining 208 customers to be analyzed.

In this data-cleaning phase, missing values of measures were filled using a neural net [29]. These failures can be due to transmission interruptions or damage in the measurement equipment. Therefore, to estimate missing values, a multi layer perceptron (MLP) artificial neural net was used.

The historical data of electricity consumption will serve as support to estimate the lacking power values, previously detected in Data Pre-processing.

The neural net was trained starting from the report of each customer's consumption. By completing this missing data, the errors of the metered load curves are attenuated without making significant alterations in the real measures. After the completion data, it has been prepared for clustering.

Each customer is represented by his representative daily load curve resulting from elaborating the data from the measurement campaign. For each customer, the representative load diagram has been built by averaging the load diagrams related to each customer [29,30]. A different representative load diagram is created to each one of the loading conditions defined: working days and weekend days. Each customer is now defined for a representative daily load curve for each of the loading conditions to be studied separately.

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

$$l^{(m)} = \left[l_1^{(m)}, \ldots l_h^{(m)} \right], \; m \in \{1 \ldots M\}, h \in \{1 \ldots H\} \qquad (5)$$

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

The diagrams were computed using the field-measurements values, and, therefore, they need to be brought together to a similar scale for the purpose of their pattern comparison. This is achieved through normalization.

For each customer the vector represented in (5) was normalized to the [0-1] range by using the peak power of its representative load diagram [29,30]. This kind of normalization was chosen to permit the maintenance of curve shape in order to compare the consumption patterns. At this point each customer is represented by a group H of data consisting of values for 15 minutes intervals which gives a set of 96 values in the range [0-1].

5.3 Determining of Electricity Customers’ Load Profiles using WEACS approach

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 the number of clusters of the final combined data partition has been fixed in 9 clusters. The expected 9 clusters were obtained for the two different load regimes: work days and weekends.

Figure 1 shows the representative load diagram obtained for each cluster using the Two-step approach and using the measurement power for the weekends.

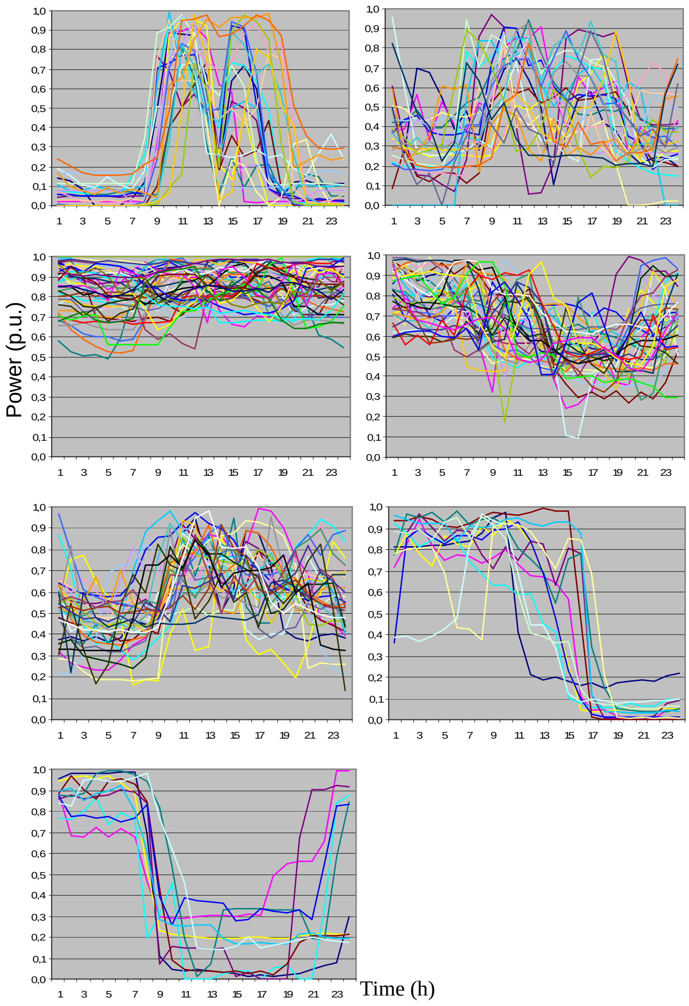

Figure 1- Clusters obtained by Two-step clustering algorithm for weekends.

Figure 2 shows the representative load diagram obtained for each cluster using the WEACS approach. and using the measurement power for the weekends. Cluster number 2 contains only six customers with atypical electric energy consumption. Their behaviors are very different of each other.

These kinds of atypical customers (outliers) should be removed from the study to avoid weakened characterization results. Apart from cluster number 2, the WEACS approach separated the customer population well and the representative load diagrams were created with a distinct load shape.

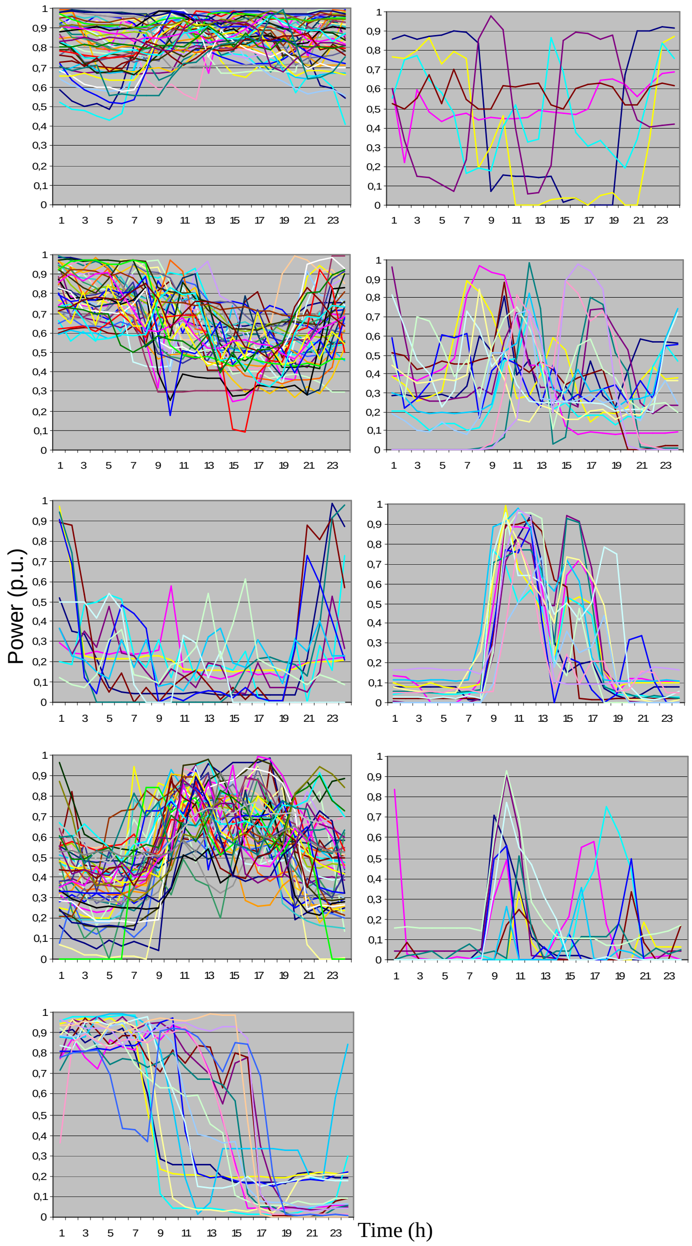

Figure 2- Clusters obtained by WEACS approach for weekends.

In order to assess the weekend clusterings results obtained by the Two-step clustering algorithm and the WEAC approach, some already referred validity indices were used: NormHubert, DB, Dunn, I, PS, Silhouette and XB. Table 2 shows the results obtained by these indices. According to the indices criteria (table 2), the data partition achieved by WEACS is considered by five of the seven indices to be better than the one obtained by Two-step. These results emphasise the conclusions that may be visually taken by analysing the graphics.

Table 2. Indices results

	Weekend		Criteria
Index	WEACS	Two-Step	
NormHub	**0.5160**	0.4957	Max
DB	**1.2343**	1.2671	Min
Dunn	0.1559	0.1778	Max
I	**0.6652**	0.6757	Max
PS	**0.4219**	0.4804	Min
Silhouette	0.3394	0.3510	Max
XB	**0.6969**	0.9594	Min

One of the representative load diagrams obtained using the WEACS approach by applying it to the measurement power for the working days also contains only seven customers with atypical electric energy consumption. These customers´ behaviors are also very different of each other and, as already mentioned, in this case this cluster should be removed from the study.

With the 8 resulting clusters, a representative load diagram has been obtained for each cluster by the load diagrams of the clients assigned to the same cluster (figure 3). The same procedure was applied to the weekend data (figure 4). Each curve represents the load profile of the corresponding customer class.

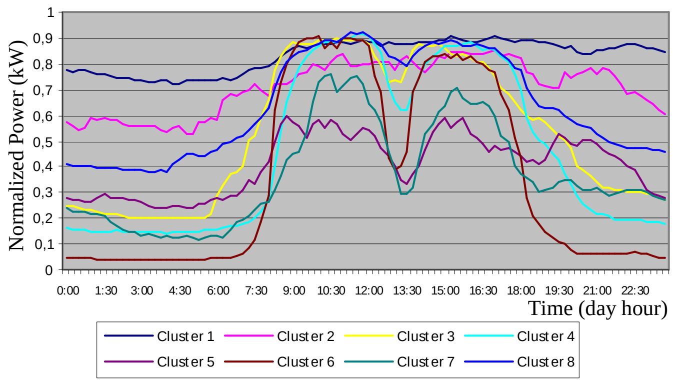

Figure 3- Representative Load Profile for working days clusters

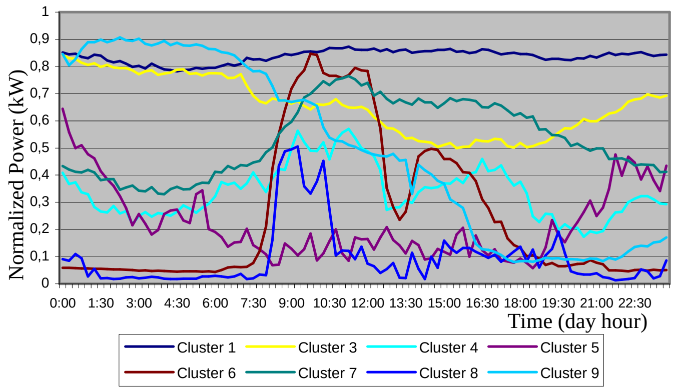

Figure 4- Representative Load Profile for weekend clusters

From the representative load diagrams obtained to each cluster it is possible to see that the WEACS approach has separated the customer population well, producing representative load diagrams with distinct load shapes.

For the characterization of the customer classes a first trial was made to search for an association between the clusters and the components of the contractual data. Specifically, an association between the activity type and the hired power of each customer and the obtained clusters was searched for. The results show that several of the customer's activity types and hired powers are present in a large number of clusters. So, a poor correlation exists between the clusters and the activity types and between the clusters and the hired power. These results show that the contractual data is highly ineffective from the viewpoint of the characterization of the customers' electrical behavior.

6. Conclusion and future work

This paper deals with the clustering of electricity customers, 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 diagrams obtained with each approach and the results of the assessment made by cluster validity indices it was noticed that the WEACS approach separates the customer population better than the Two-step cluster algorithm.

The results obtained point out that the contractual parameters are poorly connected to the load profiles, so further work is required in order to produce global shape indices able to capture relevant information on the customers' consuming behavior.

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 customer class, according to the new rules introduced in the liberalized electricity market.

Following the classification of the customers into classes, a decision support system will be developed for assisting managers in properly fixing contract details for each customer classes. This system must be sufficiently flexible to follow the variations in the customers' 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. Fred, A. K. Jain, "Evidence accumulation clustering based on the k-means algorithm," in S.S.S.P.R, T.Caelli et al., editor,., Vol. LNCS 2396, Springer-Verlag, 2002, pp. 442 –451

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

[7] A. Strehl and J. Ghosh. "Cluster ensembles - a knowledge reuse framework for combining multiple partitions", Journal of Machine Learning Research 3, 2002.

[8] S.T. Hadjitodorov, L. I. Kuncheva, L. P. Todorova, Moderate Diversity for Better Cluster Ensembles, Information Fusion, 2005.

[9] 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.

[10] 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.

[11] 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.

[12]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)
[13]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).
[14]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.
[15]M. Halkidi, Y. Batistakis, M. Vazirgiannis, "Clustering algorithms and validity measures", Tutorial paper in the proceedings of the SSDBM 2001 Conference.
[16]Theodorodis, S., Koutroubas, K., Pattern Recognition. Academic Press, 1999.
[17]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.
[18]Dunn, J.C., "Well separated clusters and optimal fuzzy partitions," J. Cybern, Vol. 4, 1974, pp. 95-104.
[19]Davies, D.L., Bouldin, D.W., "A cluster separation measure,". IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 1, No2, 1979.
[20]S.C. Sharma, Applied Multivariate Techniques, John Willwy & Sons, 1996.
[21]Calinski, R.B.& Harabasz, J, "A dendrite method for cluster analysis," Communications in statistics 3, 1974, pp.1-27.
[22]Kaufman, L. & Roussesseeuw, P., Finding groups in data: an introduction to cluster analysis, New York, Wiley, 1990.
[23]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.
[24].Xie, X.L., Beni, G., "A Validity Measure for Fuzzy Clustering," IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 13, 1991, pp. 841-847.
[25]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.
[26]J.A. Hartigan, "Statistical theory in clustering", J. Classification, 1985, 63-76.
[27]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.
[28]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.
[29]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", in Proc. of the International Conference on Knowledge Engineering and Decision Support (ICKEDS´06), pp 107-115, Lisbon, Portugal, May, 2006.
[30]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.