An Electric Energy Consumer Characterization Framework Based on Data Mining Techniques

Vera Figueiredo, Fátima Rodrigues, Zita Vale, *Member, IEEE*, and Joaquim Borges Gouveia

***Abstract*—This paper presents an electricity consumer characterization framework based on a knowledge discovery in databases (KDD) procedure, supported by data mining (DM) techniques, applied on the different stages of the process. The core of this framework is a data mining model based on a combination of unsupervised and supervised learning techniques. Two main modules compose this framework: the load profiling module and the classification module. The load profiling module creates a set of consumer classes using a clustering operation and the representative load profiles for each class. The classification module uses this knowledge to build a classification model able to assign different consumers to the existing classes. The quality of this framework is illustrated with a case study concerning a real database of LV consumers from the Portuguese distribution company.**

***Index Terms*—Classification, clustering, consumer classes, data mining, decision trees, load profiles, neural networks.**

I. Introduction

ONE of the major consequences of electricity markets liberalization is the freedom that all customers will have on the choice of their electricity supplier. This new scenario creates an environment were several retail companies compete for the electricity supply of end users. An overview about electricity retail markets is presented in [1]. To obtain well functioning markets it is essential to define new rules and structures concerning data collection and description and the definition of communication protocols between the different participants in the market. This new structures will increase significantly the amounts of data collected by the participants in the market. This data grows up in a dynamic form and can play an important role in the decision support and in the definition of market strategic behavior. The development of frameworks and tools, able to extract useful knowledge from this data, can be a competitive advantage for the participants in the market.

The knowledge of how and when consumers use electricity is essential to the competitive retail companies. This kind of knowledge can be found in historical data of the consumers collected in load research projects developed in many countries. One of the important tools defined in these projects are different consumers classes represented by its load profiles. Load profiling has been a matter of research during the last years [2]–[5]. In [2] data mining techniques are used on the determination of load profiles for different type of consumers considering the effect of weather conditions. In [3] and [4] statistical and clustering techniques are used on the determination of load profiles to support the development of tariff offer and market strategies. In [5] a knowledge discovery in databases (KDD) process, to extract useful knowledge from electricity consumers' data, is described. In this process data mining techniques are applied, on different stages of the process, to find the different consumption patterns. These patterns are represented by their load profiles and each of the patterns represents a consumer class. This knowledge is useful to develop a decision support system to support the definition of adequate contract options and market strategies.

However, the increase of data available and its dynamic growth will bring new challenges to the load profiling and consumer characterization. The new tools must be able to treat large amounts of data with all the problems common to real databases, like noise, missing values and outliers. These tools must be flexible, robust and able to provide an easy actualization of the knowledge as new data is available. This paper presents a framework developed to support the retail and distribution companies on the extraction of knowledge from electricity consumption data. This is based on the application of data mining techniques to the determination and characterization of a set of load profiles, representing the different consumption patterns of a sample of consumers. A classification tool to support the attribution of new consumers to the existing classes complements this characterization. The framework is able to treat different data sets in an easy and efficient way and provides results like consumer classes, represented by its load profiles, and classification models. These results can be updated as new data is collected.

The paper is organized as follows: In Section II the consumers' characterization framework is described. In Section III the data mining model structure and the techniques used are presented. Section IV describes a case study, using real historical data, from the Portuguese Distribution Company, to illustrate the quality of the framework. In the last section some concluding remarks are presented.

II. Customer Characterization Framework

This framework is based on a KDD [6] procedure supported by data mining (DM) techniques, applied on the different stages

Manuscript received February 13, 2004; revised October 20, 2004. This work was supported in part by the DaMICE Project supported by the Portuguese Science and Technology Foundation (FCT). Paper no. TPWRS-00073-2004.

V. Figueiredo and Z. Vale are with the Department of Electrical Engineering, Polytechnic Institute of Porto (ISEP/IPP), Porto, Portugal and GECAD (Knowledge Engineering and Decision Support Research Group) (e-mail: veraf@dee.isep.ipp.pt; zav@dee.isep.ipp.pt).

F. Rodrigues is with the Department of Computer Engineering, Polytechnic Institute of Porto (ISEP/IPP), Porto, Portugal and GECAD (e-mail: fr@dei.isep.ipp.pt).

J. B. Gouveia is with the Department of Engineering and Industrial Management, University of Aveiro (UA), Aveiro, Portugal (e-mail: bgouveia@egi.ua.pt).

Digital Object Identifier 10.1109/TPWRS.2005.846234

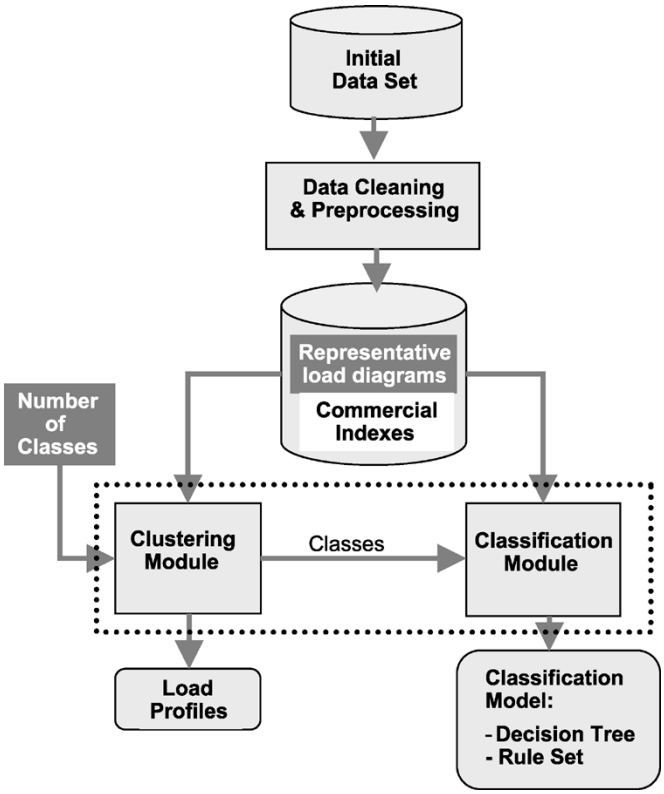

Fig. 1. Structure of the customer characterization framework.

of the process. A description of the framework's structure is presented in Fig. 1.

The following major steps can be identified.

1) *Data Selection*: the initial database, of significant dimension, stores data related with electricity consumers. The first step is the selection of the data with more significance to the process. This selection is made according to the voltage level of the consumers. Separate studies must be conducted to different voltage levels.
2) *Data Cleaning and Preprocessing*: In the cleaning phase we check for inconsistencies in the data and outliers are removed using the following procedure. Anomalous consumption values and outages are detected and replaced based on the information of similar days. In the preprocessing phase missing values are detected and replaced using regression techniques. Linear regression is used to estimate numerical attributes like missing values of measures, and logistic regression is used to estimate nominal attributes like the missing commercial information, such as activity type, tariff type, etc. With this procedure the major problems of real databases are minimized and the initial data set is cleaned and completed.
3) *Data Reduction*: The data reduction is made using previous knowledge about the way the loading conditions, like the season of the year and the type of weekday (working days or weekends), affect electricity consumption. Data are separated, according to the different loading conditions, in smaller data sets. We have two data sets representing each season of the year, one for working days and another for weekends. To obtain a more effective data reduction, without loosing important information, the data from each individual consumer are reduced. This is based on the reduction of the measured daily load diagrams, corresponding to each loading condition, to one representative load diagram. This representative load diagrams are obtained elaborating the data from the measurement campaign. For each consumer, the representative load diagram is built by averaging the measured load diagrams. Each consumer is then described by one single representative load diagram in each data set, for the different loading conditions. The diagrams are computed using the field-measurements values, so they need to be brought together to a similar scale for the purpose of their pattern comparison. This is achieved through normalization. For each consumer the vector of the representative load diagram was normalized to the [0–1] range using the peak power of the representative load diagram. This kind of normalization allows maintaining the shape of the curve and comparing the consumption patterns.
4) *Data Mining*: This step involves selection and application of the data mining techniques. This is made using one isolated technique or combining several techniques, to build a model able to find relevant knowledge about the different consumption patterns found in the data. The implementation of the model involves several steps, like attribute selection, fitting the models to the data and evaluating the models. This will be described in the next section.
5) *Interpretation of the discovered knowledge*: in this step the knowledge discovered by the DM model is improved. This knowledge can provide new insights into relationships between data elements and facilitate more productive and sophisticated decision support applications.

III. Data Mining Model

The data-mining model for consumer characterization is based on the combination of unsupervised and supervised learning techniques. After the data preprocessing and reduction phase, each consumer is represented by its representative load diagram and the commercial indexes used by the distribution company. The representative daily load diagram of the mth consumer is the vector $\mathrm{l}^{(\mathrm{m})} = \{\mathrm{l}_1^{(\mathrm{m})}, \ldots, \mathrm{l}_\mathrm{h}^{(\mathrm{m})}, \ldots, \mathrm{l}_\mathrm{H}^{(\mathrm{m})}\}$ where $\mathrm{l}_\mathrm{h}^{(\mathrm{m})}$ are the normalized values of the instant power consumed in the instant h and $\mathrm{h} = 1, \ldots, \mathrm{H}$ with $\mathrm{H} = 96$, representing the 15-min interval between the collected measurements. The commercial indexes available are of contractual nature (i.e., activity type, contracted power, tariff type, supply voltage level). The distribution company, to classify its clients, defines these indexes a priori. The proposed model is described in Fig. 1 and can be divided in two main modules according to the techniques used and to the results obtained. In the first module unsupervised learning, based on clustering techniques, is used

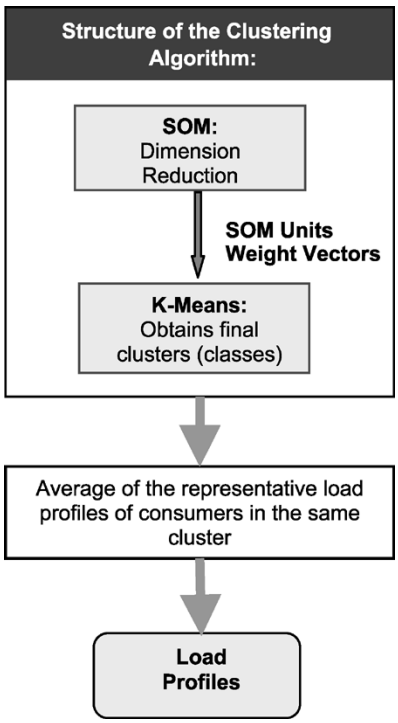

Fig. 2. Structure of the load profiling module.

to obtain a partition of the initial sample into a set of consumer classes. These classes represent the different consumption patterns existing among the sample in study. Each of these classes is represented by its load profile. In the second module, supervised learning (using decision trees) is used to describe each class by a rule set and create a classification model able to assign consumers to the existing classes. This model is important to the determination and actualization of the load profiles and the classification model as new data are collected and to the attribution of new consumers to the existing classes.

A. Load Profiling Module

The load profiling module's goal is the partition of the initial data sample in a set of classes defined according to the load shape of the representative load diagrams of each consumer. Fig. 2 presents the structure of this module. This is made assigning to the same class consumers with the most similar behavior, and to different classes consumers with dissimilar behavior. The first step of the module development was the selection of the most suitable attributes to be used by the clustering model. To obtain the best separation between the classes it is important to use the most detailed information about the shape of the consumers' load diagrams. The vectors with the normalized representative load diagrams are the best option. The selection of the most suitable clustering algorithm is described in [7] and was based on a comparative analysis of the performance of different algorithms. Several algorithms were tested performing different clustering operations. To evaluate the performance of the different algorithms two measures of adequacy were used: a measure of cluster compactness (MIA) and another measure of cluster separation (CDI) presented in [4]. The best results are obtained with a combination of a self-organizing map (SOM) [8] with the classical k-means algorithm [9]. This combination operates in two levels. In the first level the SOM is used to obtain a reduction of the dimension of the initial data set. The SOM performs the projection of the H-dimensional space, containing the M vectors representing the load diagrams of the consumers in the initial data set, into a bidimensional space. Two coordinates, representing the SOM attributes in the bidimensional space, are assigned to each client. At the end of the first level the initial data set is reduced to the number of winning units in the output layer of the SOM, represented by its weight vectors. This set of vectors is able to keep the characteristics of the initial data set and achieve a reduction of its dimension. In the second level the k-means algorithm is used to group the weight vectors of the SOM's units and the final clusters are obtained. The use of the k-means in the second level allows the definition of the number of classes as an input of the model. This combination is very interesting for large data sets, very common in data mining problems. The SOM has good performance with large data sets and is able to process large amounts of data, reducing this data to a smaller data set. During the comparative analysis it was possible to conclude that the k-means algorithm presented a very good performance with data sets with continuous attributes, like the ones we are using, but this algorithm presents limitations with large data sets. The combination of both algorithms was able to solve these limitations and create a solution able to deal with large data sets. Testing both solutions we were able to conclude that the results obtained were similar, which proves the effectiveness of the proposed combination. The load profiles for each class are obtained by averaging the representative load diagrams of the consumers assigned to the same class (cluster).

We also solved the load patterns using conventional statistic regression methods, based on the partition of the data set according to the commercial indexes. Standard deviation values obtained for each class were much lower using our methodology instead of using the classical statistic regression methods (considering the same number of classes). This stresses the validity of the proposed methodology.

B. Classification Module

The major goals of the classification module (see Fig. 3) are the following:

1) inference of a rule set to characterize each class;
2) support the attribution of new consumers to the classes obtained by the load profiling module.

The first attempt made and detailed in [5] was the search for the correlation between the commercial indexes and the classes obtained. The results point out that a poor correlation exists, so it is not possible to create a good classification model based only on the commercial indexes. This means that new indexes, able to capture relevant information about the consumption behavior, must be derived to obtain a more complete and useful consumer characterization and create the classification model. These indexes must contain information about the daily load curve shape of each consumer. Several indexes were proposed in [10] and we selected the most relevant, presented in Table I.

The classification model uses supervised learning, based on the knowledge about the relation between the characteristics

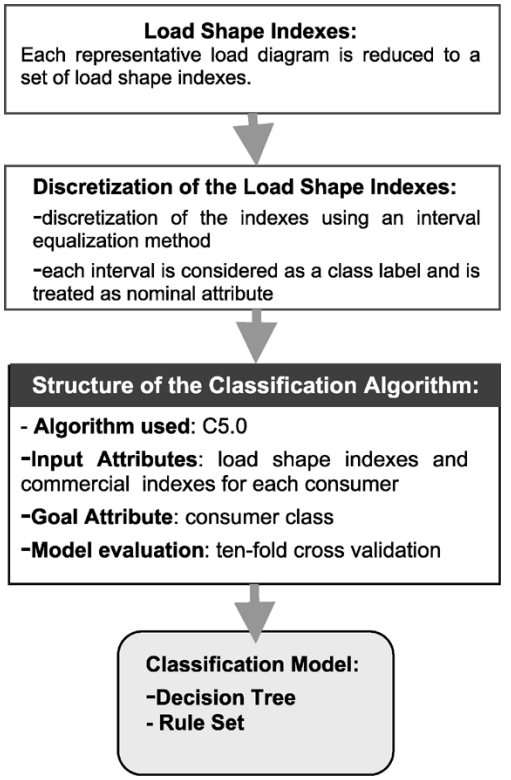

Fig. 3. Structure of the classification module.

TABLE I
NORMALIZED LOAD SHAPE INDEXES

Parameter	Definition	Period of definition
Load Factor	$d_1 = \frac{P_{av,day}}{P_{\max,day}}$	1 day
Night Impact	$d_3 = \frac{1}{3}\frac{P_{av,night}}{P_{av,day}}$	1 day (8 hours night, from 11 p.m. to 7 a.m.)
Lunch impact	$d_5 = \frac{1}{8}\frac{P_{av,lunch}}{P_{av,day}}$	1 day (3 hours from 12:00 to 15:00)

of the consumer and its corresponding class, obtained with the clustering operation.

The model's goal attribute is the consumer class obtained by the clustering module. The load shape indexes are computed, for each consumer, using the representative load diagrams. In order to obtain a reduction of the range of values assumed by these indexes, and treat them as nominal attributes, they are replaced by a small number of distinct categories using an interval equalization method [11].

This method obtains intervals with different sizes, choosing them so that approximately the same number of consumers falls in each one, to minimize the loss of information due to the replacement of the indexes by a set of discrete categories. Each interval is a class label. This allows us to treat the load shape indexes in the same manner we treat the commercial indexes. The classification model inputs are the commercial and the load shape indexes for each consumer.

The classification algorithm used is the C5.0 [12]. This algorithm was selected because it creates robust models and does not require long training times to estimate so it presents good performances with large data sets as the ones used in data mining. A divide and conquer strategy is used and the algorithm works by splitting the sample, based on the attribute that provides the maximum information gain. Each subsample defined by the first split is split again, usually based on a different attribute, and the process is repeated until the subsamples cannot be split any further. Finally, the lowest level splits are reexamined, and those that do not contribute significantly to the value of the model are removed or pruned. C5.0 can produce two kinds of models, a decision tree and a rule set. The decision tree is a straightforward description of the splits found by the algorithm. The rule set represents a simplified version of the information found in the decision tree.

The model evaluation is performed using ten-fold cross validation [11]. This kind of evaluation was selected to improve the results obtained in the presence of small data sets. Using this evaluation technique it is possible to train the algorithm using the entire data set and obtain a more precise model. This will increase the computational effort but improves the model's capacity of generalization to different data sets. The evaluation is performed by randomly splitting the initial sample in ten subsamples. The model is trained using 9/10 of the data set and tested with the 1/10 left. This is performed 10 times on different training sets and finally the ten error estimates are averaged to yield an overall error estimate.

The classification model creates a complete characterization of consumers' classes based on the most relevant attributes selected by the model. This model can be used to assign new consumers to the existing classes.

IV. CASE STUDY ON A DATA BASE OF ELECTRICITY CONSUMERS

A case study concerning a database with information from 165 LV consumers is considered in this section. This information has been gathered by measurement campaigns carried out by EDP—Distribuição, the Portuguese Distribution Company. The measurement campaigns were made during a period of three months in summer and another three months in winter for working days and weekends in each customer of the sample population. The instant power consumption for each consumer was collected with a cadence of 15 min., by real-time meters. The commercial indexes related with the activity code and the contracted power are also available. In Tables II and III it is possible to analyze the distribution of the sample population according to the commercial indexes. This data set is applied to the customer characterization framework presented in the previous sections.

A. Data Preprocessing

Data's usual bad quality is one of the major problems of working with real databases. The supplied database presented some problems including wrong information and missing data.

TABLE II
DESCRIPTION OF THE CONSUMER DATA SET STUDIED (CONTRACTED POWER)

Contracted Power (kW)	1,1	3,3	6,6	9,9	13,2	16,5	19,8	39,6	Missing values
Consumers Distribution (%)	4,6	28,7	21,3	23,0	7,5	1,7	6,3	1,2	5,8

TABLE III
DESCRIPTION OF THE CONSUMER DATA SET STUDIED (ACTIVITY TYPE)

Activity Type	A	B	C	E	G	H	Missing values
Consumers Distribution (%)	61,49	17,24	1,72	4,60	11,5	2,30	1,2

A – Domestic; B – Non domestic; C –Buildings Illumination; E – Industrial; G – Public Illumination; H – Others

Due to the limited dimension of the sample we had to minimize the loss of data so the cleaning and preprocessing phase assumed major importance. In the data-cleaning phase wrong information has been corrected. The outliers and outages were detected using interactive graphics and regression techniques based on the data from similar days, which permits replacement by most probable values.

The missing values of measures were detected and estimated using regression techniques. With these data cleaning and completion procedures, the quality of the data has been improved with a minimum loss of information. These data were reduced and normalized using the procedures presented in Section II. A different representative load diagram is created and normalized to each one of the loading conditions defined: winter, summer, working days and weekends.

In this phase each consumer is described by a normalized representative daily load curve for each of the loading conditions to be studied separately. In this section, we present the results obtained for the winter working days and weekend data sets, to illustrate our case study.

B. Definition of the Number of Classes

The number of classes is an input of the model so it must be defined based on a criterion that leads to an adequate selection. The number of classes obtained by the clustering module must be in the range 2 to $\sqrt{M}$, where M is the number of consumers in the data set. Based on information from the electricity company, we fixed a minimum number of 6 and a maximum number of 9 classes. To define the number of classes, several clustering operations were performed to study the evolution of the clusters compactness using the measure Mean Index Adequacy (MIA) presented in [4]. The following distances (1) and (2) are defined to assist the formulation of the adequacy measure.

1) Distance between two load diagrams

$$d(li,lj)=\sqrt{\frac{1}{H}\times\sum_{h=1}^{H}(li(h)-lj(h))^2}. \tag{1}$$

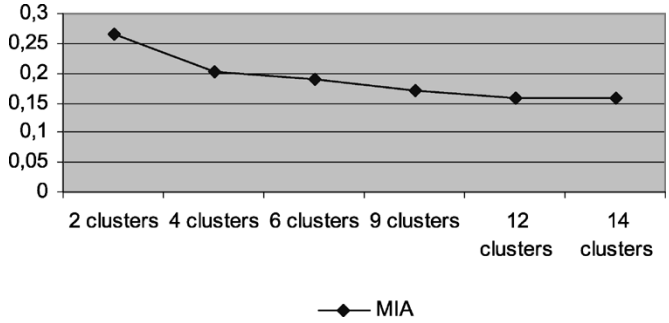

Fig. 4. MIA evolution with the number of clusters.

2) Distance between a representative load diagram and the center of a set of diagrams

$$d\left(r^{(k)},L^{(k)}\right)=\sqrt{\frac{1}{n^{(k)}}\sum_{m=1}^{n^{(k)}}d^2\left(r^{(k)},l^{(m)}\right)}. \tag{2}$$

Let us consider a set of M load diagrams separated in k classes with $k=1,\ldots,K$, were K is the total number of clusters, and each class is formed by a subset $C^{(k)}$ of load diagrams, where $r^{(k)}$ is a pattern assigned to cluster k. The MIA is defined by

$$MIA=\sqrt{\frac{1}{K}\sum_{k=1}^{K}d^2\left(r^{(k)},C^{(k)}\right)}. \tag{3}$$

The smaller values of MIA indicate more compact clusters. The k-means algorithm was used to study the cluster tendency of the data set based on the MIA measure. The obtained results are presented in Fig. 4. It is possible to see that 9 clusters would be the best choice, considering the indication of the distribution company and the evolution of the MIA, because for more then 9 clusters the improvement on the clusters compactness, represented by the decrease of the MIA values, is not very relevant.

C. Consumer Characterization

A different consumer characterization is obtained to each of the data sets corresponding to the different loading conditions considered. Each data set was applied to the module and a complete characterization was obtained, represented by a load profile and a rule set describing each class. The winter-working days and winter-weekends data sets are applied to the clustering module. In the first level a SOM is trained to obtain the initial reduction of the data sets. A rectangular grid with dimension 7×10 is used. The SOM has the following architecture: Input layer: 96 units and Output layer: 70 units. The winning units vectors in the output layer represent the reduced data set to be clustered by the k-means, in the second level. The final number of clusters is introduced as an input in this level. Table IV presents the final distribution of the consumers through the 9 clusters. Each cluster corresponds to a different class. The algorithm isolated the consumers with atypical behavior in clusters with small number of elements. The clusters representing atypical behavior

TABLE IV
NUMBER OF CONSUMERS WITHIN EACH CLUSTER

	Cluster	1	2	3	4	5	6	7	8	9
Number of Consumers	Working Days	17	1	13	60	6	1	37	3	29
	Weekends	16	20	7	1	33	36	19	30	2

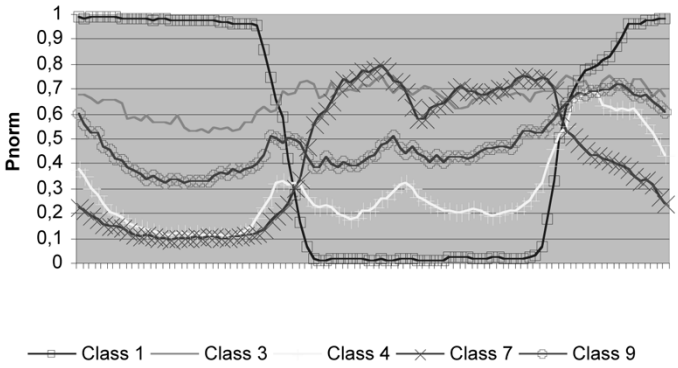

Fig. 5. Load profiles of winter working day classes.

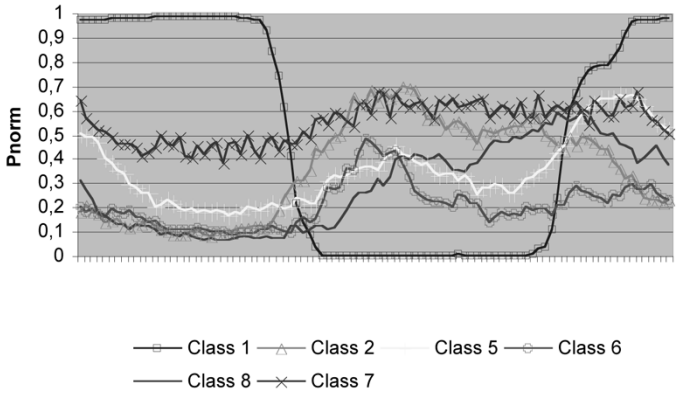

Fig. 6. Load profiles of winter weekend classes.

do not represent typical consumption patterns so they will not be considered to build the classification model.

The load profiles of the typical behavior classes are obtained by averaging the representative load diagrams of the consumers in the same cluster. The load profiles representing the winter-working day classes are presented in Fig. 5 and the load profiles for weekend classes are represented in Fig. 6. Analyzing the load diagrams profiles obtained it is possible to see that the clustering module has well separated the client population, creating representative load diagrams with distinct load shapes. Each cluster was also validated calculating its standard deviation, and the results obtained were quite satisfactory.

The representative load diagrams are used by the classification module to compute the load shape indexes d1, d3 and d5 of each consumer. The attributes representing each consumer, used by the classification algorithm, are described in Table V.

The classification module is able to produce classification models represented by a decision tree and a rule set, based on the most relevant attributes. These are the attributes selected by the model to grow the decision tree.

TABLE V
ATTRIBUTES USED BY THE CLASSIFICATION ALGORITHM

Name	Type	Description
Contracted Power (CP)	nominal	{1,1; 3,3; 6,6;9,9;13,2;16,5;19.0; 33; 39,6}kW
Activity Type	nominal	{ A, B, C, E, G}
Load Factor (d1)	nominal	<0,2; 0,2-0,3; 0,3-0,4; 0,4-0,5; 0,5-0,6; >0,6
Night impact (d3)	nominal	0-0,2; 0,2-0,3; 0,3-0,4; 0,4-0,6; >0,6
Lunch impact (d5)	nominal	0-0,1; 0,1-0,15; 0,15-0,2; >0,2
Class (cluster)	nominal	{cluster1,…,cluster9}

TABLE VI
CHARACTERISTICS OF THE CLASSIFICATION MODELS

Data Set	Overall Accuracy	Relevant Attributes	Rule Set
Working Days	81%	d1 and d3	13 Rules
Weekends	74%	d1, d3 and CP	15 Rules

```
if d1 ∈ [0.4-0.5[ and d3∈ [0.4-0.6[ then class 1
if d1 ∈ [0.4-0.5[ and d3 >= 0.6 then class 1
if d1 ∈ [0.5-0.6[ and d3∈ [0.4-0.6[ then class 1
if d1 ∈ [0.5-0.6[ and d3 >= 0.6 then class 1
if d1 ∈ [0.5-0.6[ and d3 ∈ [0.3-0.4[ then class 3
if d1 >= 0.6 then class 3
if d1 ∈ [0.2-0.3[ then class 4
if d1 ∈ [0.3-0.4[ then class 4
if d1 ∈ [0.4-0.5[ and d3 ∈ [0-0.2[ then class 7
if d1 ∈ [0.5-0.6[ and d3 ∈ [0.2-0.3[ then class 7
if d1 ∈ [0.5-0.6[ and d3 ∈ [0-0.2[ then class 7
if d1 ∈ [0.4-0.5[ and d3 ∈ [0.2-0.3[ then class 9
if d1 ∈ [0.4-0.5[ and d3 ∈ [0.3-0.4[ then class 9
```

Fig. 7. Rule set for the winter working days classification model.

The data from the consumers in the typical classes were used to build the models for winter working days and weekends. These were evaluated using ten-fold cross validation and the overall accuracy obtained is presented in Table VI. The characteristics of the models obtained are also presented in Table VI.

The model, according to the data set, selected different relevant attributes. It is possible to conclude that the load factor and the night impact are the most relevant attributes to describe the consumers' characteristics. Fig. 7 presents an example with the rule set obtained for the winter working days data set, for this case study. The obtained rules are simple and with straightforward interpretation. These rules can be integrated in a decision support system.

V. CONCLUSION

This paper deals with the characterization of electricity consumers using historical data. A new and robust framework, based on DM techniques, to find relevant knowledge about how and when consumers use electricity, is proposed. The DM model presented is able to obtain a set of consumer classes, represented by its load profiles, and a classification model represented by a rule set. The innovative contributions of this

framework are the ability to treat large data sets, its robustness in the presence of missing data and outliers and the combination of unsupervised and supervised learning to perform a more complete characterization of the classes obtained and create a classification tool to support the practical application of these classes.

The proposed framework is generally applicable to different databases and can be used to update the consumer characterization as new data are collected. The quality of this framework is illustrated by a case study considering a real database, supplied by the Portuguese Distribution Company. The results obtained were quite satisfactory considering the limitations of the available database. This can be a useful tool to the distribution and retail companies to support the definition and selection of the most adequate electricity supply contracts to suit each of its client's needs. The development of these kinds of applications is of crucial importance nowadays, due to the pressure of competitive retail markets and the increase of the amount of data available, with the evolution of the data collected and stored.

ACKNOWLEDGMENT

The authors express their gratitude to EDP Distribuição, the Portuguese Distribution Company, for supplying the data resulting from the load research project carried out during the nineties, which has been used to test the proposed framework and for the support given in different phases of this work.

REFERENCES

[1] E. Hirst and B. Kirby, *Retail Load Participation in Competitive Wholesale Electricity Markets*. Washington, DC: Edison Electric Institute Rep., 2001.

[2] B. Pitt and D. Kirchen, "Applications of data mining techniques to load profiling," in *Proc. IEEE PICA*, Santa Clara, CA, May 1999, pp. 131–136.

[3] C. S. Chen, J. C. Hwang, and C. W. Huang, "Application of load survey to proper tariff design," *IEEE Trans. Power Syst.*, vol. 12, no. 4, pp. 1746–1751, Nov. 1997.

[4] G. Chicco, R. Napoli, P. Postulache, 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.

[5] V. Figueiredo, F. J. Duarte, F. Rodrigues, Z. Vale, and J. Gouveia *et al.*, "Electric energy customer characterization by clustering," in *Proc. ISAP*, Lemnos, Greece, Sep. 2003.

[6] U. Fayyad, G. Piatetsky-Shapiro, P. J. Smith, and R. Uthurasamy, "From data mining to knowledge discovery: an overview," in *Advances in Knowledge Discovery and Data Mining*. Cambridge, MA: AAAI/MIT Press, 1996, pp. 1–34.

[7] F. Rodrigues, V. Figueiredo, F. J. Duarte, and Z. Vale, "A comparative analysis of clustering algorithms applied to load profiling," in *Lecture Notes in Artificial Intelligence (LNAI 2734)*. New York: Springer-Verlag, 2003, pp. 73–85.

[8] T. Kohonen, *Self-Organization and Associative Memory*, 3rd ed. Berlin: Springer-Verlag, 1989.

[9] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[10] M. Ernoult and F. Meslier, "Analysis and forecast of electrical energy demand," *Revue Général d'Electricité*, no. 4, 1982.

[11] I. Witthen and E. Frank, *Data Mining—Practical Machine Learning Tools and Techniques With Java Implementations*. New York and San Mateo, CA: Morgan Kaufmann Publishers, Academic Press, 2000.

[12] R. Quinlan, *The Book C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1993.

Vera Figueiredo received the B.Sc. degree in 1999 and the M.Sc. degree in 2003 from the University of Porto, Porto, Portugal.

She is currently an Assistant Professor of Electric Power Systems with the Polytechnic Institute of Porto, Porto, Portugal. Her research interests include competitive electricity markets, load forecasting, consumer characterization, and data mining.

Fátima Rodrigues received the B.Sc. degree from the University of Minho, Minho, Portugal, in 1989, the M.Sc. degree from the University of Porto, Porto, Portugal, in 1997, and the Ph.D degree in computer science from the University of Minho in 2000.

She is currently a Coordinator Professor of Computer Engineering with the Polytechnic Institute of Porto, Porto, Portugal. Her research areas include data mining, KDD, and artificial intelligence.

Zita Vale (M'86) received the B.Sc. degree in 1986 and the Ph.D. degree in electrical engineering in 1993 from the University of Porto, Porto, Portugal.

She is currently a Coordinator Professor with the Polytechnic Institute of Porto, Porto, Portugal. Her research areas include power systems operation and control, electricity markets, decision support, and artificial intelligence.

Joaquim Borges Gouveiareceived the B.Sc. degree in 1973 and the Ph.D. degree in electrical engineering in 1983 from the University of Porto, Porto, Portugal.

He is currently a Full Professor with the University of Aveiro, Aveiro, Portugal. His research areas include energy efficiency, innovation, and services operation management.