CHARACTERIZATION OF MV CONSUMERS USING HIERARCHICAL CLUSTERING

S. Ramos[1], V. Figueiredo[1], F. Rodrigues[2], R. Pinheiro[1], Z. Vale[1]

[1] Department of Electrical Engineering, Polytechnic Institute of Oporto, Portugal
GECAD – Knowledge Engineering and Decision Support Group
{sramos, veraf, raul, zav}@dee.isep.ipp.pt
[2]Department of Computer Engineering, Polytechnic Institute of Oporto, Portugal
GECAD – Knowledge Engineering and Decision Support Group
fr@dei.isep.ipp.pt

Abstract – With the electricity market liberalization, distribution and retail companies are looking for better market strategies based on adequate information upon the consumption patterns of its electricity customers. A fair insight on the customers' behaviour will permit the definition of specific contract aspects based on the different consumption patterns. In this paper, we propose a Knowledge Discovery in Databases project applied to electricity consumption data from a utility client's database. To form the different customers' classes, and find a set of representative consumption patterns, we have used the Two-Step algorithm which is a hierarchical clustering algorithm. Each consumer class will be represented by its load profile resulting from the clustering operation. Next, to characterize each consumer class a classification model will be constructed with the C5.0 classification algorithm.

Keywords: electricity markets, load profiles, hierarchical clustering, classification.

I. INTRODUCTION

In the past, 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. In deregulated electricity markets the knowledge about customer's consumption patterns (daily load profile), is extremely important for the accomplishment of agreements in the price of the electricity between consumers 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 costumers 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 [1]. One possible method of differentiation is the development of tailored contracts defined according to customer consumption patterns.

In 2004, Portugal adheres to the Iberian Market of Electricity (MIBEL) jointly with Spain, what will open up new opportunities of business to electricity companies, namely for those who sell electricity to customers of Medium Voltage (MT) and Low Voltage (LV), in a deregulated environment. Agents acting in this market operate in any area of the country or in Iberian Peninsula and can establish bilateral contracts or make proposals of purchase or sale in the spot market.

The installation of real time measurement equipment will simplify the load forecast task. On an open electricity market, all consumers (MV and LV) should have appropriate metering equipment and the metering services should be assured by an independent market agent. If for the LV customers the installation of that equipment in all of them will be an exhausting task requesting long time and huge investments, in the MV customers the installation of these measurement equipments is already mandatory. So, in the case of MT consumers, the sellers of electricity can use the incoming data from the measurement equipment to reduce the risk in the purchase of energy to the producers.

One of the important tools defined using this data are the load profiles for different consumers 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. The accurate classification of consumer classes and the association of a load profile are essential to support marketing strategies.

This paper is organized as follows: in section 2 we briefly resume the hierarchical clustering algorithm and the cluster method used. In section 3 the C5.0 classification algorithm is described. In 4 we present normalized shape indicators for characterizing the load profiles. In section 5 we present our case study – a sample of medium voltage consumers from the Portuguese Distribution Company and finally, in last section, some conclusions and future work are presented.

II. CLUSTERING ALGORITHMS

When trying to discover knowledge from data, one of the first arising tasks is to identify groups of similar objects that are to carry out cluster analysis for obtaining data partitions. There are several clustering methods that can be used for cluster analysis. Yet for a given data set, each clustering method may identify groups whose member objects are different. Thus a decision must be taken for choosing the clustering method that

produces the best data partition for a given data collection. In order to support such a decision, we have used indices for measuring the quality of the data partition. To choose the optimal clustering schema, we will follow two proposed criteria:

- compactness: members of each cluster should be as close to each other as possible;
- separation: the clusters should be widely spaced from each other.

A. *Hierarchical Clustering Algorithms*

A hierarchical algorithm yields a *dendrogram* representing the nested grouping of patterns and similarity levels at which groupings change. It leaves considerable flexibility in implementation. The dendrogram can be broken at different levels to yield different partitions of the data set. Most hierarchical clustering algorithms are variants of the single-link [2], complete-link [3], and minimum-variance [4] algorithms. Of these, the single-link and complete-link algorithms are most popular. These two algorithms differ in the way they characterize the similarity between a pair of clusters. In the single-link method, the distance between two clusters is the *minimum* of the distances between all pairs of patterns drawn from the two clusters (one pattern from the first cluster, the other from the second). In the complete-link algorithm, the distance between two clusters is the *maximum* of all pair wise distances between patterns in the two clusters. In either case, two clusters are merged to form a larger cluster based on minimum distance criteria. The complete-link algorithm produces tightly bound or compact clusters [5]. The single-link algorithm, by contrast, suffers from a chaining effect [6]. It has a tendency to produce clusters that are straggly or elongated. The clusters obtained by the complete-link algorithm are more compact than those obtained by the single-link algorithm. The single-link algorithm is more versatile than the complete-link algorithm, otherwise. However, from a pragmatic viewpoint, it has been observed that the complete-link algorithm produces more useful hierarchies in many applications than the single-link algorithm [7].

B. *Two-Step Algorithm*

This cluster method is a scalable cluster analysis algorithm designed to handle very large data sets. It can handle continuous and categorical variables or attributes. It requires only one data pass. It 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.

The pre-cluster step uses a sequential clustering approach. It scans the data records one by one and decides if the current record should be merged with the previously formed clusters or starts a new cluster based on the distance criterion. The procedure is implemented by constructing a modified cluster feature (CF) tree. The CF-tree consists of levels of nodes, and each node contains a number of entries. A leaf entry represents a final sub-cluster. If the CF-tree grows beyond allowed maximum size, the CF-tree is rebuilt based on the existing CF-tree by increasing the threshold distance criterion. The rebuilt CF-tree is smaller and hence has space for new input records. This process continues until a complete data pass is finished [8]. All records falling in the same entry can be collectively represented by the entry's CF. When a new record is added to an entry, the new CF can be computed from this new record and the old CF without knowing the individual records in the entry. Note that the CF-tree may depend on the input order of the cases or records. To minimize the order effect, it is necessary randomly order the cases.

The cluster step takes sub-clusters resulting from the pre-cluster step as input and then groups them into the desired number of clusters using an agglomerative hierarchical clustering method [9]. In general, the larger the number of sub-clusters produced by the pre-cluster step, the more accurate the final result is. However, too many sub-clusters will slow down the clustering during the second step.

A distance measure is needed in both the pre-cluster and cluster steps. Two distance measures are available, the log-likelihood distance measure, can handle both continuous and categorical variables, and the Euclidean distance only applied if all variables are continuous.

All experiments that will be described in the following sections were conducted using Clementine version 7.1. This is an integrated DM toolkit, which uses a visual-programming interface, and supports all KDD stages.

III. Classification Algorithms

In classification problems a set of pre-classified data points are given and the classification algorithm tries to discover a rule, which allows mimicking as closely as possible the observed classification. A classification problem is a supervised learning task where the output information is a discrete classification.

A. *C5.0 Algorithm*

The C5.0 algorithm [10] works by splitting the sample based on the field that provides the maximum information gain. Each sub sample defined by the first split is then split again, usually based on a different field, and the process repeats until the sub samples cannot be split any further. Finally, the lowest level splits are re-examined, 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 is a straightforward description of the splits found by the algorithm. Each terminal, or "leaf" node describes a particular subset of the training data, and each case in the training data belongs to exactly one terminal node in the tree. So, only one prediction is possible for any particular data record presented to a decision tree.

In contrast, a rule set is a set of rules that tries to make predictions for individual records. Rule sets are derived from decision trees and, in a way, represent a simplified or distilled version of the information found in the decision tree. Rule sets can often retain most of the important information from a full decision tree but with a less complex model. Because of the way rule sets work, they do not have the same properties as decision trees. The most important difference is that with a rule set, more than one rule may apply for any particular record, or no rules at all may apply. If multiple rules apply, each rule gets a weighted "vote" based on the confidence associated with that rule, and the final prediction is decided by combining the weighted votes of all of the rules that apply to the record in question. If no rule applies, a default prediction is assigned to the record.

C5.0 models are quite robust in the presence of problems such as missing data and large number of input fields. They usually do not require long training times to estimate. In addition, C5.0 models tend to be easier to understand than some other model types, since the rules derived from the model have a very straightforward interpretation. C5.0 also offers a powerful boosting method to increase accuracy of classification.

IV. Indices to Characterize the Electricity Consumers

A. Commercial Indices

These indices are related to the electricity contracts and are defined a priori by the electricity distribution company to obtain a classification of its clients. They give us information about each consumer's activity type, hired power, tariff option and supply voltage level. These type of indices have no relation with the load diagram, so if used isolated they can't provide a good consumer classification. As this is the only information obtained from the contractual data it is important to include it in the classification process.

B. Normalized Shape Indicators

These indices are derived from the daily load diagrams and some of them are based on the set of indices proposed in [11]. They give information about the daily load curve shape and so about the consumption pattern of each consumer. The indices can be used as attributes in the classification process to introduce this kind of information and to obtain a more accurate consumer classification, or by other hand, In Table 1 is presented a set of indices where Pmax is the maximum power demand, Pmin is the minimum power demand and Pav is the average power demand during a representative day.

Table I- Normalised shape indicators for characterising the load profiles

Parameter	Definition	Acquisition Period
Daily P_{av}/P_{max}	$fc = \frac{P_{av,day}}{P_{max,day}}$	1 day
Daily P_{min}/P_{max}	$fv = \frac{P_{\min,day}}{P_{\max,day}}$	1 day
Daily P_{min}/P_{av}	$mm = \frac{P_{min,day}}{P_{av,day}}$	1 day
Night Impact	$in = \frac{1}{3}\frac{P_{av,night}}{P_{av,day}}$	1 day (8 hours night, from 11 p.m. to 6 a.m.)
Daily P_{av}/P_{inst}	$fu = \frac{P_{av,day}}{P_{inst}}$	1 day
Lunch Impact	$ia = \frac{1}{8}\frac{P_{av,lunch}}{P_{av,day}}$	1 day (2 hours lunch, from 12 a.m. to 2 p.m., 16 hours daytime from 6 a.m. to 23 p.m.)

V. Case Study

Broadly speaking, the KDD process consists of three phases, namely pre-processing, data mining and post-processing [12].

These phases can be overlapped and the results produced in previous iterations can be used to improve the following iterations of the process. The pre-processing phase includes three broad sub-phases, namely data selection, data cleaning and data transformation (not necessarily in this order).

A. Data Selection

Our case study is based on a set X = 208 MV consumers from a Portuguese utility. Information on the customer consumption has been gathered by measurement campaigns carried out by EDP Distribuição - Portuguese Distribution Company, between 1995/96, and these data was used for the purpose of a study demonstration. These campaigns were based on a load research project where the previous definition of a sample population, type of consumers (MV, LV), where meters were installed, sampling cadence (15, 30 minutes...) and total duration (months, years...) of data collection were defined. The instant power consumption for each customer was collected with a cadence of 15 minutes, which gives 96 values a day for each client, for each day. The measurement campaigns were made during a period of 3 months in summer and another 3 months in winter for working days and weekends of the a sample population of MV consumers. There is also available

for this population the commercial data related with the monthly energy consumption, the activity code and the hired power.

B. Data Pre-processing

There are always problems with data. That explains why previous to any DM process it is indispensable a data-cleaning phase to detect and correct bad data, and a data-treatment phase to derive data accordingly to DM algorithms that will be used [13]. In the data-cleaning phase we have filled missing values of measures by a neural net. These failures can be due to transmission interruptions or damage in the measurement equipment.

Therefore, to estimate missing values we used a multi layer perceptron (MLP) artificial neural net.

With the information given by the measurements equipment, we made several time fields used for training the neural net, namely:

$$Day(year) = \frac{quarter.hour.year}{96} \quad (1)$$

$$Hour(year) = \frac{quarter.hour.year}{4} \quad (2)$$

$$Week = \frac{Day.year}{7} \quad (3)$$

$$Hour(day) = \frac{[quarter.hour.year - (Day.year \times 96)]}{4} \quad (4)$$

Starting from the report of each customer's consumption, the neural net was trained. We can verify the consumption estimates in figure 1:

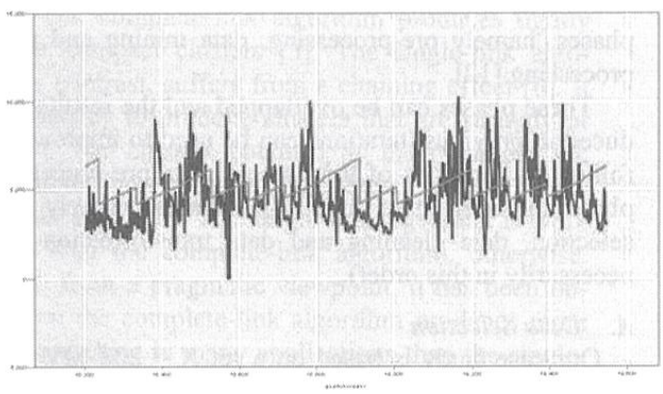


Figure 1- Estimation of a MV consumer consumption with a neural net.

With this data completion the errors of the metered load curves are attenuated without making big changes in the real measures.

After this data completion we prepare data for clustering. Each customer is represented by its 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 [14]. A different representative load diagram is created to each one of the loading conditions defined: annual week days and annual weekend days. Each customer is now defined for a representative daily load curve for each of the loading conditions to be studied separately. We present the study performed for the period: annual weekends.

The representative daily load diagram of the m^{th} consumer is the vector l(m) = {lh(m)} with h = 1,...,H where H = 96, representing the 15 minutes intervals in a day.

The diagrams were 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 l(m) was normalized to the [0-1] range by using the peak power of its representative load diagram. This kind of normalization permits maintaining the shape of the curve 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].

C. Data Mining Operations

1) Clustering

A clustering procedure based on Two-step algorithm has been used to group the load patterns on the basis of their distinguishing features. This algorithm was selected based on a comparative analysis of the performance of different clustering algorithms applied to this data set presented in [14]. If the number of clusters is unknown the clustering can be repeated for a set of different values between 2 to $\sqrt{X}$, where X is the number of patterns in the pattern set. As the goal of our clustering is finding a set of load profiles to characterize the consumers and, in the future, to study tariff offers, the number of clusters must be small enough to allow the definition of different tariff structures to each class. Based on information from the electricity company we fixed a minimum number of 6 and a maximum number of 9 clusters.

We performed different clustering exercises using different numbers of clusters, to evaluate the evolution of the indexes MIA (Mean Index Adequacy) and CDI (Clustering Dispersion Indicator) with the number of clusters [11].

The Mean Index Adequacy (MIA) depends on the average of the mean distances between each pattern assigned to the cluster and its center.

$$MIA = \sqrt{\frac{1}{K}\sum_{k=1}^{K} d^2(r^{(k)}, C^{(k)})} \quad (5)$$

The Clustering Dispersion Indicator (CDI) depends on the distance between the load diagrams in the same cluster and (inversely) on the distance between the class representative load diagrams.

In (6) R is the set of the class representative load diagrams.

$$CDI = \frac{\sqrt{\frac{1}{K}\sum_{k=1}^{K}\left[\frac{1}{2.n^{(k)}}\sum_{n=1}^{n^{(k)}} d^2(l^{(m)}, C^{(k)})\right]}}{\sqrt{\frac{1}{2K}\sum_{k=1}^{K} d^2(r^{(k)}, R)}} \quad (6)$$

The clustering algorithm that produces the smaller MIA and CDI values prevails over the others in term of performance.

As it can be seen from figure 2, we can conclude that the indexes decrease as the number of clusters increases, and for a superior number of 9 clusters the reduction gain it is not very significantly.

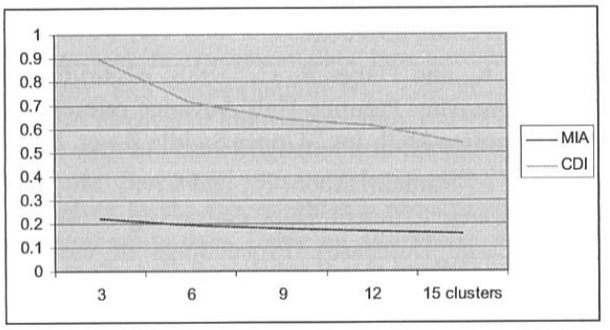
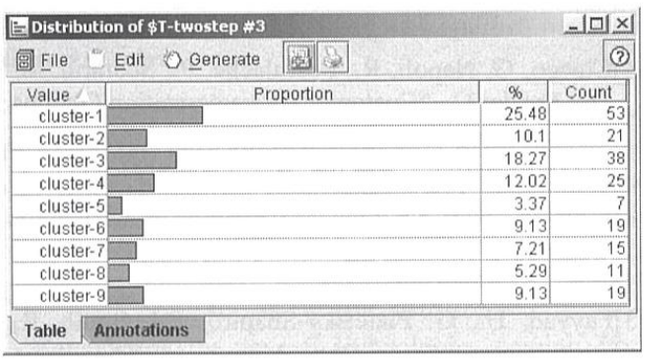


Figure 2- Indexes (MIA and CDI) evolution with the number of clusters.

We used the representative load diagrams, dealing with the normalized shape indicators to obtain the clusters. The Two-step algorithm was applied to obtain the expected 9 clusters (Figure 3).



Value	Proportion	%	Count
cluster-1		25.48	53
cluster-2		10.1	21
cluster-3		18.27	38
cluster-4		12.02	25
cluster-5		3.37	7
cluster-6		9.13	19
cluster-7		7.21	15
cluster-8		5.29	11
cluster-9		9.13	19

Figure 3- Final clusters obtained with the Two-step algorithmic (using the normalized indicators to represent the load diagrams)

With the resulting clusters obtained with Two-Step algorithm we obtained the representative diagram for each cluster for weekends and week days for a period of one year, averaging the load diagrams of the clients assigned to the same cluster. The next figure shows the representative load diagram obtained for each cluster.

Each curve represents the load profile of the corresponding costumer class.

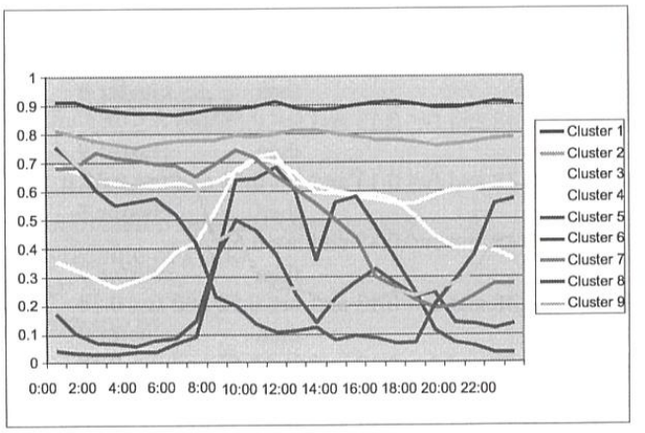


Figure 4- Representative Load Profile for each customer class

In [15], a trial was performed to search for associations between the clusters and the components of the contractual data (commercial indices). The results have showed that a poor correlation exists between the main clusters and the contractual data. These results have proved that the contractual data is highly ineffective from the viewpoint of the characterization of the electrical behavior of the costumers.

B. Classification

To obtain more relevant information to describe the consumption patterns of each cluster population we have used a rule-based modelling technique, the C5.0 classification algorithm. In this first phase, we have chosen this algorithm due it is easier to understand since the rules derived from the model have a very straightforward interpretation. It is our aim to use other algorithms and compare the results obtained with C5.0.

The C5.0 algorithm has analyzed the 208 representative daily load diagrams of the consumers and we used a neural net for analysing the importance of each attribute (indices). We realised that the "fu" indictor (fu=Pav/Pinst) only had an importance of 1,2%, therefore it was removed from the classification.

We obtained the importance classifications for each indicator using an MLP neural net, described in table 2, where the factors are the presented in section IV.

Table II- Relative Importance of Input

fv	0.27154
in	0.189893
ia	0.181249
mm	0.177296
fc	0.102986

Figure 5 presents an example of a rule set obtained from the C5.0 algorithm for the weekend's data set, which is relevant for this case study. The obtained rules are simple and of straightforward interpretation. It is possible to conclude that the nonuniformity coefficient, the low factor, the night impact, the lunch impact and load factor are relevant attributes to describe the con-

sumers' characteristics.

if mm ≤ 0.48 and fv≤ 0.13 and ia≤ 0.55 and fc≤ 0.35 and in≤ 0.31	then	**cluster 8**
if mm ≤ 0.48 and fv≤ 0.13 and ia≤ 0.55 and fc≤ 0.35 and in >0.31	then	**cluster 9**
if mm ≤ 0.48 and fv≤ 0.13 and ia≤ 0.55 and fc> 0.35	then	**cluster 5**
if mm ≤ 0.48 and fv≤ 0.13 and ia> 0.55	then	**cluster 7**
if mm ≤ 0.48 and fv> 0.13 and ia≤ 0.52 and fv≤ 0.18	then	**cluster 5**
if mm ≤ 0.48 and fv> 0.13 and ia≤ 0.52 and fv> 0.18	then	**cluster 4**
if mm ≤ 0.48 and fv> 0.13 and ia> 0.52	then	**cluster 4**
if mm > 0.48 and fv≤ 0.45 and	then	**cluster 3**
if mm > 0.48 and fv> 0.45 and fv≤ 0.64	then	**cluster 3**
if mm > 0.48 and fv> 0.45 and fv> 0.64	then	**cluster 1**

Figure 5- Rule set for the Weekend day's classification model.

The model has been tested and it's overall accuracy was approximately of 90% (89,4% to be more exactly), which shows that the results are reasonably satisfactory.

VI. Conclusion and Further Work

This paper deals with the clustering of the electricity consumers based on the normalized indicators to create the representative load diagrams. The Two-step clustering algorithm was used and the clusters characterization is performed using C5.0 classification algorithm.

The results point out that the contractual parameters are poorly connected to the load profiles.

The clustering algorithm was able to produce load profiles with distinctly different load shapes and the classification algorithm presents a good overall accuracy. The rules obtained are simple and with straightforward interpretation.

It is our aim to compare the efficiency of the algorithm C.5.0 with different classification algorithms.

It is also our aim to develop a decision support system for assisting the managers in properly fixing the best tariff structure for each customer class. This one must be sufficiently flexible to follow the variations in the load patterns of the customers.

Acknowledgment

The authors express their gratitude to EDP Distribuição, the Portuguese Distribution Company, for supplying the data used in this work and for the support given in different phases of this work.

References

[1] Grønli, H., Livik, K., Pentzen, H., "Actively Influencing on Customer Actions – Important Strategic Issues for the Future., *DA/DSM Europe Distrib-uTECH Conference*, Amsterdam 14 – 16 October 1998, Proceedings.

[2] Sneath, P.H. and Sokal R.R. 1973. Numerical Taxonomy. Freeman, London, UK.

[3] King B. 1967. Step-wise clustering procedures. J.Am. Stat. Assoc. 69, 86-1001.

[4] Murtagh F. 1984. A survey of recent advances in hierarchical clustering algorithms which use cluster centers. Comput. J. 26, 354-359.

[5] Baeza-Yates R. A. 1992. Introduction to data structures and algorithms related to information retrieval. In Information Retrieval Data Structures and Algorithms, W. B. Frakes and R. Baeza-Yates, Eds. Prentice-Hall, Inc., Upper Saddle River, NJ, 13-27.

[6] Nagy G. 1968.State of the art in pattern recognition. Proc. IEEE 56, 836-862.

[7] Jain A. K. and Dubes R. C. 1988. Algorithms for clustering data. Prentice-Hall Advance reference series. Prentice-Hall, Inc., Upper Saddle River, NJ.

[8] Zhang, T., Ramakrishnon, R., and Livny M. (1996). BIRCH: An Efficient Data Clustering Method for Very Large Datebases. Proceedings of the ACM SIGMOD Conference on Management of Data, p. 103-114, Montreal, Canada.

[9] Chiu, T., Fang, D., Chen, J., Wang, Y., and Jeris, C. (2001). 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.

[10]Quinlan 1993 The book, C4.5: Programs for Machine Learning. *Morgan Kaufmann Publishers*

[11]Chicco, G, Napoli, R., Postulache, P., Scutariu, M. And Toader C., "Customer Characterization Options for Improving the Tariff Offer", *IEEE Transactions on Power Systems*, Vol. 18, N°1, February 2003.

[12]Frawley, W.J., G. Piatetsky-Shapiro, C. Matheus, "Knowledge Discovery in Databases: An Overview", *AI Magazine*, 1992.

[13]Fayyad, U., G. Piatetsky-Shapiro, P.J. Smith, R. Uthurasamy, "From Data Mining to Knowledge Discovery: An Overview". *In Advances in Knowledge Discovery and Data Mining*, pages 1-34. AAAI/MIT Press, 1996.

[14]Rodrigues F, Duarte F.J, Figueiredo V., Vale Z., Cordeiro M.., 2003, "A Comparative Analysis of Clustering Algorithms Applied to Load Profiling.", *Proceedings of MLDM 2003*, Leipzig, Germany.

[15]Figueiredo V., Duarte F.J.,Rodrigues F., Vale Z., Ramos, C. Ramos, S., Gouveia B., 2003, "Electric Energy Customer Characterization by Clustering.", *Proceedings of ISAP 2003*, Lemnos, Greece.