Electric Energy Customer Characterization by Clustering


Abstract-- With the electricity market liberalization, the distribution and retail companies are looking for better market strategies based on adequate information upon the consumption patterns of its electricity customers. A fair insight on the customers’ behavior will permit the definition of specific contract aspects based on the different consumption patterns.

The knowledge about how and when consumers use the electricity has an important role in a free and competitive electricity market, but this one grows up in a dynamic form. The treatment of this data must be made with the application of Data Mining and Knowledge Discovery techniques to support the development of generic load profiles to each consumer’s class. In this paper, we propose a KDD project applied to electricity consumption data from an utility clients data base. To form the different customers classes a comparative analysis of the performance of the Kohonen Self Organized Maps (SOM) and K-means algorithm for clustering is presented. Each customer class is then represented by its load profile.

Index Terms-- electricity markets, clustering, self-organized-maps, K-means, load profiles, KDD, Data Mining.

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 will create an unstable environment where all customers have access to the market. The electricity suppliers will need strategies for approaching customers based on cost leadership or differentiation. 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 the deregulated market, 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.

To study the electricity consumption patterns the development of load research projects is essential. In a load research project a sample population is defined, real time meters are installed in these clients and the data collected in these meters during the period in study is treated and analyzed to create tools to support companies. Load research projects were developed in many countries [2], [3]. One of the important tools defined in these projects 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 to the class are useful to:

- proper design of marketing strategy and rate tariff;
- accurate load forecasting for better system expansion planning to achieve good service quality;
- proper design of energy conservation and load management strategies to improve system operation efficiency;
- replace the real time meters to support market settlement.

In [4] we can see how the definition of load profiles can support the development of dedicated market strategies. The definition of load profiles assumes major importance in small consumers where the use of real time meters is not economically interesting.

To move from profiling in abstract to profiling in practice a number of decisions is required and the most important and critical are the number and coverage of the profiles. Each load profile should represent a relatively homogeneous group of consumers and must be distinctly different from the others. The number of profiles should be big enough to represent the different consumption patterns but small enough to be used on the definition of market strategies.

This paper is organized as follows: in section 2 we present an overview of KDD process followed by the description of K-means and Kohonen Networks clustering methods. In section 4 we present a case study using data from a sample of low voltage consumers from the Portuguese Distribution Company and finally, in last section some conclusions and future work are presented.

II. KDD - AN OVERVIEW

Knowledge Discovery in Data Bases (KDD) is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [5].

KDD is an inter-disciplinary subject, formed by the
intersection of many different areas, such as, machine learning, pattern recognition, databases, statistics, artificial intelligence, and data visualization. KDD systems typically draw upon methods, algorithms, and techniques from these diverse fields. The unifying goal is the extraction of knowledge from data in the context of large databases. The KDD process is interactive and iterative, involving numerous steps with many decisions being made by the user, as described in figure 1 [6].

![Fig. 1. KDD Process](image)

Broadly speaking, the KDD process consists of three phases, namely pre-processing, data mining and post-processing. 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 cleaning, data selection, and data transformation (not necessarily in this order). Note that some operations of these sub-phases can be partially performed during the construction of a data warehouse, but it is often desirable to perform some additional pre-processing which is particularly suitable for KDD purposes (or even for a specific DM algorithm).

Data cleaning may involve consistency checking, data-error correction, filling in or eliminating null values, removing outliers, etc. Data cleaning methods are inherently application-domain dependent, and the participation of the end-user is crucial. The user usually makes the decision about whether or not that data should be actually considered as noise (and so be removed from database for KDD purposes), by using his domain knowledge. In the context of mining very large databases, a major kind of data selection is the selection of relevant attributes from the large set of database attributes.

Pre-processing is followed by the Data mining (DM) phase, where the algorithms chosen are applied to the pre-processed data. DM involves fitting models to, or extracting patterns from data without the additional steps of the KDD process (such as incorporating appropriate prior knowledge and interpretation of the results). Different methodologies can be used, depending if DM goals are: discover hidden associations; predictable sequences; or accurate classifications. It should be noticed that this phase is estimated to take only about 20% of the overall effort of the entire KDD process [5]. However, this phase is by far the most studied in the KDD literature, probably due to the fact that the automation of the other phases is an even more difficult task.

Finally the output of the DM algorithm can be refined in the post-processing phase. This phase may involve interpretation of the discovery knowledge or some post-processing of this knowledge. In general, a major goal of this phase is to improve the comprehensibility of the knowledge discovered by the DM algorithm. This knowledge can provide new insights into relationships between data elements and facilitate more productive and sophisticated decision support applications.

Our first goal was to group clients in well-defined and non-overlapping classes represented by its load profile. This will give us an insight on the way electricity consumption evolves during the day. The load profiles can be defined according to different load conditions, like the day of the week and the season of the year. In our work we define load profiles to support the definition of different tariffs to each costumer class according to the new rules introduced in the liberalized electricity market. In the next sections we will describe in detail the load profiling DM case study developed.

### III. CLUSTERING ALGORITHMS

#### A. Kohonen Networks

Kohonen Networks [9] are unsupervised topologically preserving artificial neuronal network that performs clustering analysis of the input data. The basic units are the neurons, and these are usually organized in a two dimensional layers: the input layer, and the output layer, which is often referred to as the output map. All the input neurons are connected to all the output neurons, and these connections have “strengths” associated with them.

Input data is presented to the input layer and the values are propagated to the output layer. Each output neuron then gives a response. The output neuron with the strongest response is said to be the winner, and is the answer for that input. Initially all weights are random. In order to train, an input pattern is shown and the winner adjusts its weights in such a way that it reacts even more strongly next time it sees that (or a very similar) record. Also, its neighbors (those neurons surrounding it) adjust their weights so they also react more positively. All the input records are shown and weights updated accordingly. This process is repeated many times until the changes being made are very small.

The most attractive feature of the SOM is that, once trained, the map represents the projection of the data set belonging to an N-dimensional space, into a bi-dimensional one.

All experiments that will be described in the following sections were conducted using Clementine version 6.5. This is an integrated DM toolkit, which uses a visual-programming interface, and supports all KDD stages [10].

#### B. K-Means

The K-means algorithm [7] is the most widely used clustering algorithm. The name comes from representing each
of the k clusters by the mean, or weighted average, of its points, the so-called cluster center. It requires the initialization of the number of clusters and starts by assigning k cluster centers randomly selected from the pattern set. Then it proceeds by assigning each pattern from the initial set to the nearest cluster center and recomputes the center using the current cluster memberships, until the convergence criterion is met. Typically the convergence criteria are: no patterns are reassigned to a new cluster center or minimal decrease square error. This algorithm has the advantage of clear geometrical and statistical meaning but works conveniently only with numerical attributes. It is also sensitive to outliers.

IV. ADEQUACY MEASURES OF CLUSTERING OUTCOMES

The load diagrams grouping is performed using the SOM as a clustering algorithm and a K-means algorithm in a comparative analysis. To assess the efficiency of each clustering process a measure of adequacy should be employed. For this goal we use the indices proposed in [11]. These indices are based in distances computed with the multi-dimensional vectors used to represent the load diagrams of each consumer and the load profile of each class.

To assist the formulation of the adequacy measures the following distances are defined:

a) Distance between two load diagrams

\[ d(li, lj) = \sqrt{\frac{1}{H} \sum_{h=1}^{H} (li(h) - lj(h))^2} \]  

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

\[ d(r^{(k)}, L^{(k)}) = \sqrt{\frac{1}{N^{(k)}} \sum_{m=1}^{N^{(k)}} d^2(r^{(k)}, l^{(m)})} \]  

Using the distances (1) and (2) its possible to define performance measures to evaluate the clustering tools. A good clustering tool is able to determine, well-separated classes of load diagrams and assure that the load diagrams assigned to the same class are very similar. Considering a set of X load diagrams separated in K classes with k= 1, ..., K and each class is formed by a subset C^{(k)} of load diagrams, where r^{(k)} is a pattern assigned to cluster k, the following performance measures are defined:

A. Mean Index Adequacy (MIA)

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

\[ MIA = \sqrt{\frac{1}{K} \sum_{k=1}^{K} d^2(r^{(k)}, C^{(k)})} \]  

B. Clustering Dispersion Indicator (CDI)

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

In (4) R is the set of the class representative load diagrams.

\[ CDI = \sqrt{\frac{1}{2K} \sum_{k=1}^{K} \sum_{m=1}^{N^{(k)}} d^2(l^{(m)}, C^{(k)}) - \frac{1}{2} \sum_{k=1}^{K} \sum_{m=1}^{N^{(k)}} d^2(r^{(k)}, R) \]  

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

V. FINDING CUSTOMER LOAD PROFILES BY CLUSTERING

We present a case study on a set X = 165 LV consumers from a Portuguese utility.

Information on the customer consumption has been gathered by measurement campaigns carried out by EDP – Distribuição, Portuguese Distribution Company. 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 in each costumer of the a sample population of LV consumers. There is also available for this population the commercial data related with the monthly energy consumption, the activity code and the hired power.

A. 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. In the data-cleaning phase we have filled missing values of measures by linear regression. There were also some values of activity type and hired power missing which were computed using logistic regression because these are classified as categorical attributes.

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 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 [4]. A different representative load diagram is created to each one of the loading conditions defined: winter, summer, working days and weekends.

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 working days and weekends during winter.

The representative daily load diagram of the m<sup>th</sup> consumer is the vector \( l^{(m)} = \{ l^{(m)}_h \} \) with \( h = 1, \ldots, H \) where \( H = 96 \) representing the 15 minutes interval.

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 and compares the consumption patterns.

At this point each customer is represented by a group \( H \) of data consisting on values for 15 minutes intervals what gives a set of 96 values in the range [0,1]

**B. Data Mining Operations**

A clustering procedure based on Kohonen Networks (SOM) [9] has been used to group the load patterns on the basis of their distinguishing features. If the number of clusters is unknown the clustering can be repeated for a set of different values between 2 and \( \sqrt{X} \), 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 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. The Kohonen network performs the projection of the \( H \) – dimensional space, containing the \( X \) vectors representing the load diagrams, into a bi-dimensional space. To each client are assigned two coordinates \( SKX \) and \( SKY \) representing the Kohonen net attributes.

To perform this clustering operation we trained a network with the following architecture:

- **Input layer:** 96 units
- **Output layer:** 16 units

When applied to this set of patterns the network performs a projection of the input values in a bi-dimensional space with 9 different sets of coordinates representing 9 different clusters (Fig 2):

In order to perform a comparative analysis, to select the most suitable clustering algorithm, we perform another clustering operation using the K-means algorithm. This algorithm requires the initialization of the number of clusters that we set to 9 to keep the number obtained with the Kohonen network.

To evaluate the two algorithms we compute the MIA and CDI indexes with the results obtained with each algorithm (Tab 1). From the results obtained we conclude that the K-means algorithm has a better performance with this data set. This can be explained because to obtain a small number of clusters the dimension of the SOM’s output grid is small and the best performances are obtained with bigger grids.

![Fig. 2. Clusters formation with the Kohonen network.](image)

![Fig. 3. Clusters obtained with k-means.](image)

<table>
<thead>
<tr>
<th>MIA AND CDI OBTAINED FOR 9 CLUSERS WITH THE KOHONEN NETWORK AND K-MEANS</th>
<th>Kohonen</th>
<th>K-means</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIA</td>
<td>0.1950</td>
<td>0.1694</td>
</tr>
<tr>
<td>CDI</td>
<td>0.9358</td>
<td>0.5890</td>
</tr>
</tbody>
</table>

We also performed different clustering exercises using different numbers of clusters, to evaluate the evolution of the indexes with the number of clusters. As it can be seen from
figure 4 we can conclude that, as expected, the indexes decrease as the number of clusters increases.

With the resulting clusters obtained with K-means we obtained the representative diagram for each cluster for working days and weekends in the winter period by averaging the load diagrams of the clients assigned to the same cluster. (Fig 5, Fig 6). The next two figures show the representative load diagram obtained for each cluster. Each curve represents the load profile of the corresponding costumer class.

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. From Tables 1 and 2 we conclude 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.

Further work is needed in order to produce global shape indices able to capture relevant information on the costumer’s consumption behavior. To obtain more relevant information to describe the consumption patterns of each cluster population we intend to use a rule-based modeling technique (C5.0 algorithm), to analyze those clusters, and to obtain their descriptions based on a set of indices derived from the daily load curves.
VI. CONCLUSION AND FURTHER WORK

This paper deals with the clustering of the electricity consumers based on their measured daily load curves. To obtain the clusters for the different consumer classes two algorithms were used: the Kohonen network and K-means algorithm. The K-means algorithm presented a better performance than the Kohonen network. Each customer class is represented by its load profile. 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.

The results obtained so far point out that the contractual parameters are poorly connected to the load profiles, and that fixing the price coefficients for each customer classes will be very difficult. The decision support system for assisting the managers in properly fixing the price coefficients for each customer classes will be developed. This one must be sufficiently flexible to follow the variations in the load patterns of the customers.

The number of customers of different activity types and their hired power within each weekend cluster is shown in Table III.

<table>
<thead>
<tr>
<th>Activity Type</th>
<th>Hired Power</th>
<th>Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1,1</td>
<td>1,2,3,4,5,6,7,8,9</td>
</tr>
<tr>
<td></td>
<td>3,3</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>6,6</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>9,9</td>
<td>-</td>
</tr>
<tr>
<td>B</td>
<td>3,3</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>6,6</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>9,9</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>13,2</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>19,8</td>
<td>-</td>
</tr>
<tr>
<td>C</td>
<td>6,6</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>9,9</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>13,2</td>
<td>-</td>
</tr>
<tr>
<td>D</td>
<td>9,9</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>13,2</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>19,8</td>
<td>-</td>
</tr>
<tr>
<td>E</td>
<td>9,9</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>13,2</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>19,8</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>39,6</td>
<td>-</td>
</tr>
<tr>
<td>F</td>
<td>3,3</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>6,6</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>9,9</td>
<td>-</td>
</tr>
</tbody>
</table>

VII. REFERENCES


[8] F. Jorge Duarte received his M.Sc degree in the FEUP in 2000. He joined the ISEP in 1993 as Coordinator Professor. His research areas include energy efficiency, electricity markets, decision support and Artificial Intelligence.

[9] F. Rodrigues received her Ph.D degree in Computer Science from the University of Minho in 2000. She joined ISEP in 1993 where she is a Coordinator Professor. Her research areas include Data Mining and Artificial Intelligence.

[10] Z. Vale received her Ph.D degree in Electrical Engineering at the University of Oporto in 1993. She joined ISEP in 1998 as Coordinator Professor. Her research areas include Data Mining, Decision Support and Artificial Intelligence.

[11] C. Ramos received his Ph.D degree in Electrical Engineering from the University of Oporto in 1993. He joined ISEP in 1993 as Coordinator Professor. His research areas include Knowledge Engineering and Decision Support System.

[12] S. Ramos graduated from the ISEP in 1999 where he is Assistant Professor since March 2001. He is currently working on his M.Sc thesis in the FEUP. His research areas include energy efficiency, electricity markets and load research.

[13] J. Gouveia received his Ph.D degree in Electrical Engineering from the University of Aveiro in 1983. He joined the University of Aveiro in 2000 as Full Professor. His research areas include energy efficiency, innovation and services operation management.

VIII. BIOGRAPHIES

Vera Figueiredo graduated in the Faculty of Engineering of University of Oporto (FEUP) in 1998. She joined ISEP in 2000 as Assistant Professor. She is currently working on her M.Sc thesis in the FEUP. Her research areas include competitive electricity markets, load research and Data Mining.

F. Jorge Duarte received his M.Sc degree in the FEUP in 2000. He joined the ISEP in 1996 as Assistant Professor. He is currently working on his Ph.D. thesis in University of Trás os Montes e Alto Douro. His research areas include competitive electricity markets, Data Mining and Artificial Intelligence (AI).

F. Rodrigues received her Ph.D degree in Computer Science from the University of Minho in 2000. She joined ISEP in 1993 where she is a Coordinator Professor. Her research areas include Data Mining, KDD and AI.

Z. Vale received her Ph.D degree in Electrical Engineering at the University of Oporto in 1993. She joined ISEP in 1998 as Coordinator Professor. Her research areas include Power Systems, Electricity Markets, Decision Support and Artificial Intelligence.

C. Ramos received his Ph.D degree in Electrical Engineering from the University of Oporto in 1993. He joined ISEP in 1993 as Coordinator Professor. His research areas include Knowledge Engineering and Decision Support System.

S. Ramos graduated from the ISEP in 1999 where he is Assistant Professor since March 2001. He is currently working on his M.Sc thesis in the FEUP. His research areas include energy efficiency, electricity markets and load research.

J. Gouveia received his Ph.D degree in Electrical Engineering from the University of Aveiro in 1983. He joined the University of Aveiro in 2000 as Full Professor. His research areas include energy efficiency, innovation and services operation management.