A Two-Level Clustering Approach Applied to 
Electricity Load Profiling

Jorge Duarte¹, Fátima Rodrigues¹, Vera Figueiredo², Zita Vale²

¹Department of Computer Engineering, Polytechnic Institute of Oporto, Portugal
GECAD – Knowledge Engineering and Decision Support Group
{jduarte, fr}@dei.isep.ipp.pt

²Department of Electrical Engineering, Polytechnic Institute of Oporto, Portugal
GECAD – Knowledge Engineering and Decision Support Group
{veraf, zav}@dee.isep.ipp.pt

Abstract. With the electricity market liberalization, the distribution and retail companies are looking for better market strategies based on adequate information upon the consumption patterns of its electricity customers. A fair insight on the customers’ behavior will permit the definition of specific contract aspects based on the different consumption patterns. In this paper, we propose a KDD 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, a comparative analysis of the performance of the K-means, Kohonen Self Organized Maps (SOM) and a Two-Level approach is made. Each customer class will be represented by its load profile obtained with the algorithm with best performance in the data set used.

1 Introduction

The major consequence 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 competitive environment will include the small low voltage (LV) consumers and new competitive retail companies will enter in the market. These companies will need strategies for approaching customers based on cost leadership or differentiation by additional value 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 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.

The consumption pattern of an electricity consumer is defined for his daily load diagram. These diagrams represent the evolution of the power consumed during the period of a day. The load diagram evolution can be affected for several load factors that influence the way each consumer uses electricity. These factors can be external, like the weather condition, the day of the week or related with the type of consumer (a
domestic consumer does not have the same pattern as an industrial consumer). The influence of these factors must be considered in any study related with the consumers’ electrical behavior. 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 are essential to support marketing strategies.

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 briefly resume each one of the clustering methods used: K-means, Self Organized Maps (SOM) and a Two-Level Approach. In next section we present our case study – a sample of low voltage consumers from a Portuguese Distribution Company, next a comparison between the different results obtained are performed, and finally, in last section, some conclusions and future work are presented.

2 Clustering Algorithms

When trying to discover knowledge from data one of the first arising tasks is to identify groups of similar objects that is 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 will use indexes 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.
2.1 K-means

The K-means algorithm [4] 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 criterions are: no patterns are reassigned to a new cluster center or minimal decrease square error is reached. This algorithm has the advantage of clear geometrical and statistical meaning, but works conveniently only with numerical attributes. It is also sensible to outliers.

2.2 Kohonen Self Organized Maps (SOM)

Kohonen Networks are a specific kind of 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 Kohonen Self Organized Maps (SOM) [5] is that, once trained, the map represents the projection of the data set belonging to an N-dimensional space into a bi-dimensional one.

2.3 Two-Level Approach

If SOM is used as a clustering algorithm each map unit define a small cluster consisting of the samples in its Voronoi set. The prototype vector of each unit represents the cluster center. Because of the neighborhood relations, neighboring prototypes are pulled to the same direction, and thus prototype vectors of neighboring units resemble each other. However, the optimal situation in the SOM is not when the number of prototypes equals the number of clusters. Instead, the number of units in the output grid must be much bigger then the expected number of clusters. It is possible, according to the definition of the neighborhood function, to drawn
neighboring units together, and thus neighbor units reflect the properties of the same cluster. The transition from one cluster to another on the map takes place over several map units. If a small number of clusters is desired, to avoid the reduction of the grid to a small one, the SOM units must be clustered [6]. The SOM can be used as an intermediate step and make the clustering a two level approach, improving the quality of the final results. In the first level the SOM implements an ordered dimensionality reduction, mapping the data set to the output grid, reducing the computational load. The prototype units of the output layer can be clustered using either a partitive or a hierarchical method. The clustering task is now easier because is performed in a smaller number of samples. In the second level the prototypes can be directly clustered or some specific features of the SOM can be used.

3 ADEQUACY MEASURES OF CLUSTERING OUTCOMES

The load diagram grouping is performed using the K-means, SOM and a Two-Level Approach 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 [7]. 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. They are adequate to perform a comparative evaluation between different algorithms used to obtain the same number of clusters. 

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

a) Distance between two load diagrams

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

(1)

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

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

(2)

Using the distances (1) and (2) it is 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 \(k = 1, \ldots, K\) 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:

Mean Index Adequacy (MIA)

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

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

Clustering Dispersion Indicator (CDI)

The Clustering Dispersion Indicator (CDI) depends on the distance between the load diagrams in the same cluster and (versely) on the distance between the class representative load diagrams. In (4) R is the set of the class representative load diagrams.

\[
CDI = \frac{1}{K} \sum_{k=1}^{K} \left[ \frac{1}{2.n^{(k)}} \sum_{m=1}^{n^{(k)}} d^2(i^{(m)}, C^{(k)}) \right]^{\frac{1}{2}}
\]

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

All experiments that will be described in the following sections were conducted using Clementine version 7.1 [8]. This is an integrated DM toolkit, which uses a visual-programming interface, and supports all KDD stages.

4 Finding Customer Load Profiles By Clustering

Broadly speaking, the KDD process consists of three phases, namely pre-processing, data mining and post-processing [9]. 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).
4.1 Data Selection

Our case study is based 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. 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 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.

4.2 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 [10]. 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 categorical attributes. With this data completion the errors of the metered load curves were 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 [11]. 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 weekends during the winter period. The representative daily load diagram of the $m^{th}$ consumer is the vector $l(m) = \{l_{h}(m)\}$ with $h = 1,...,H$ 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 to compare the consumption patterns.

At this point each customer is represented by a group $H$ of data consisting on values for 15 minutes intervals which gives a set of 96 values in the range $[0,1]$. 
4.3 Data Mining Operations

A clustering procedure based on Kohonen Networks (SOM) 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 and 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. 1).

![Fig. 1 Clusters formation with the Kohonen network](image)

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 same number obtained with the Kohonen network. (Fig 2).
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 grid has a number of units close to the expected number of clusters. To obtain best performances the number of units in the output layer must be much bigger than the expected number of clusters. After several experiments we conclude that is not possible to obtain 9 clusters with a number of units bigger then 16, as used in the previous approach. To improve these results we experimented a Two-Level approach. In the first level the SOM was used to reduce the dimension of the data set. In the second level the prototype vectors of the SOM units are clustered using K-means to obtain 9 clusters. The SOM used in the first level has the following architecture: Input layer - 96 units; Output layer - 70 units (rectangular grid 10×7). When applied to this data set the network performs a projection of the input values in a bi-dimensional space with 55 different sets of coordinates (Fig. 3) representing a reduction of the initial data set from the dimension 165 to 55 instances.
In the second level, the K-means algorithm is used to perform the clustering of the SOM units and obtain the expected 9 clusters (Fig. 4).

![Clusters formation with the Two-Level Approach](image)

After the second level we have the initial data set grouped into 9 different clusters, which can be used to create the different load profiles.

To evaluate the different algorithms we compute the MIA and CDI indexes with the results obtained with each algorithm (Table 1). From the results obtained we conclude that the Two-Level Approach has a better performance with this data set.

<table>
<thead>
<tr>
<th></th>
<th>K-means</th>
<th>SOM</th>
<th>Two-Level Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIA</td>
<td>0.1694</td>
<td>0.1950</td>
<td>0.1622</td>
</tr>
<tr>
<td>CDI</td>
<td>0.5890</td>
<td>0.9358</td>
<td>0.5860</td>
</tr>
</tbody>
</table>

A significant improvement has been obtained using the Two-Level Approach when compared with the performance of the SOM. This can be explained by the increase of the output layer used in this approach that is much bigger than the expected number of clusters. However, when compared with K-means the better performance of this approach is not so significant because the data set used, mainly composed by numeric attributes, is adequate to K-means algorithm.

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 5 we can conclude that, as expected, the indexes decrease as the number of clusters increases.
Fig. 5. MIA and CDI evolution with the number of clusters

With the resulting clusters obtained with Two-Level Approach we obtained the representative diagram for each cluster for weekends in the winter period by averaging the load diagrams of the clients assigned to the same cluster. (Fig 6, Fig 7). The next two figures show the representative load diagram obtained for each cluster. Each curve represents the load profile of the corresponding customer class.

Fig. 6. Representative Load Profile for consumers with non typical behavior
From the representative load diagrams obtained to each cluster it is possible to see that the clustering method used has well separated the client population, creating representative load diagrams with distinct load shapes. The method used has successfully isolated non-typical representative load diagrams what is proved by the load profiles represented in figure 6.

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 figures 8 and 9 we conclude that a poor correlation exists between the main clusters and the contractual data. These results show 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 customer’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 [12]), to analyze those clusters, and to obtain their descriptions based on a set of indices derived from the daily load curves.
5 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 three algorithms were used: K-means, SOM and a Two-Level Approach. The K-means algorithm presented a better performance than the Kohonen network. However, the Two-Level Approach presents the best performance, which indicates this combination is a good practice to improve the performance of SOM as a clustering algorithm.

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

As further work we intend to extend the attributes space with contractual data, such as, categorical attributes in order to obtain load profiles related with contractual data. After the division of the customers into classes, this will be achieved by the characterization of each cluster.

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.

References