USE OF DATA MINING TECHNIQUES TO CHARACTERIZE MV CONSUMERS AND TO SUPPORT THE CONSUMER-SUPPLIER RELATIONSHIP

SÉRGIO RAMOS[1], ZITA VALE[1], JOÃO SANTANA[2] and FÁTIMA RODRIGUES[3]

[1] Department of Electrical Engineering, Polytechnic Institute of Porto, Portugal
GECAD – Knowledge Engineering and Decision Support Group
[2]Department of Electrical Engineering, Technical Superior Institute, Lisbon, Portugal
[3]Department of Computer Engineering, Polytechnic Institute of Porto, Portugal
GECAD – Knowledge Engineering and Decision Support Group
PORTUGAL

Abstract: - This paper consists in the characterization of electric power profiles of medium voltage (MV) consumers, based on the data base knowledge discovery process. Data Mining techniques were used as support for the agents of the electric power retail markets, with the purpose of obtaining specific knowledge of their customers' consumption habits. A hierarchical clustering algorithm is used in order to form the different customers' classes and to find a set of representative consumption patterns. A classification model was built that, when applied to new consumers, allows classifying them in one of the obtained classes. New tariff options were also defined, taking into consideration the typical consumption profile of the class to which the customers belong. With the results obtained, the consequences that these will have in the interaction between customer and electric power suppliers are analyzed.

Key-Words: - classification, clustering, data mining, electricity markets, load profiles, new tariff structures.

1 Introduction

In the last decades, the electric sector has suffered deep changes due to the restructuring of the electric power markets. The end of regulated monopolies and the introduction of free competition promoted the emergence of new market agents.

In the past, the industrial history was always told from the perspective of the producer and never from the perspective of the consumer [1]. However, with today's global economy and the liberalization of numerous economy sectors, including the electrical sector, consumers have a high influence on the design of these new markets. Gradually, consumers have been treated as a demand side capable of making rational decisions, instead of a simple load that needs to be served.

In the traditional utility environment, the information about the customer's consumption was important for managing the power demand, system planning, or definition of better tariffs. The liberalization process brought not only more freedom in trading, but also the freedom of choice of the supplier by the electric power consumers.

With the emergence of competitive electricity markets, a reduced price paid by customers for electricity consumption is expected. For the retail companies, the knowledge about customer's consumption patterns will be extremely important for the accomplishment of agreements in the price of electricity between consumers and suppliers, the definition of marketing policies and innovative contracts and services.

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 [2]. One possible method of differentiation is the development of tailored contracts defined according to customer consumption patterns.

The characterization and definition of customer's classes (clusters) can be conveniently extracted by the knowledge of the real costumers' electrical behavior and also by additional external features information, such as weather data, type of activity, contracted power value, consumed energy and tariff type. In this paper we aim to characterize the representative load diagrams of MV customers, using a given number of daily load diagrams extracted from a monitoring campaign, and also develop new tariff structures based on the pattern of the representative load diagrams. For the characterization of MV consumers load profile, hierarchical clustering algorithm was used.

This paper is organized in the following manner: in section 2, we briefly summarize the hierarchical clustering algorithm and the cluster method used. In section 3, we present our case study – a sample of MV consumers from the Portuguese Distribution Company. In section 4, we develop new tariff structures based on the consumer's characterization. In section 5, we present some advantages of the typical profile and in which way it can improve the relationship between consumers and the electric power supplier. Finally, in the 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, 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.

2.1 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 [3], complete-link [4], and minimum-variance [5] 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. 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 [6]. The single-link algorithm, by contrast, suffers from a chaining effect [7]. 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 [8].

2.2 Two-Step Algorithm

This clustering 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.

A complete explanation of this clustering method can be found in [9], [10] and [20].

3 Case Study

All experiments that will be described in the following sections were conducted using Clementine version 8.5 [Clementine Data Mining System, web page – http://www.spss.com]. This is an integrated DM toolkit, which uses a visual-programming interface, and supports all Knowledge Discovery in Databases (KDD) stages.

The KDD process consists of three phases, namely pre-processing, data mining and post-processing [13].

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). Figure 1 represents our study process structure which includes several phases, starting from the data acquisition, data treatment, application of data mining algorithms, consumers' characterization, new consumers' classification, development of new tariff structures, and the analysis of the obtained results in the relationships between consumer and electricity supplier.

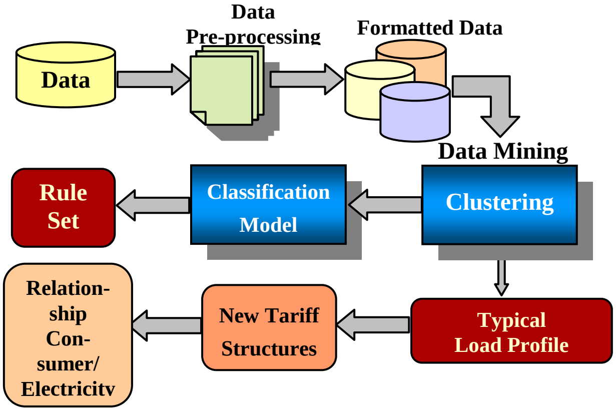


Figure 1- Study Process Structure

3.1 Data Selection

Our case study is based on a set of 229 MV consumers from a Portuguese utility service. Information on the customer consumption was gathered by measurement campaigns carried out by the Portuguese Distribution Company in the 1990's. For this population, there was also the commercial data related to the monthly energy consumption, the activity code, and the contracted power.

3.2 Data Pre-processing

There are always problems with data which is why a previous data-cleaning phase to detect and correct bad data is indispensable to any DM process, as well as a data-treatment phase to derive data according to DM algorithms that will be used [14]. In the data-cleaning phase, we have detected some damaged files and some customers without registered values, which were removed from the initial data sample. However, in this data-cleaning phase, we filled missing values of measures using a neural net [20]. 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.

The historical data of electricity consumption will serve as support to estimate the lacking power values, previously detected in Data Pre-processing. The neural net was trained starting from the report of each customer's consumption. We can verify the consumption estimates in figure 2:

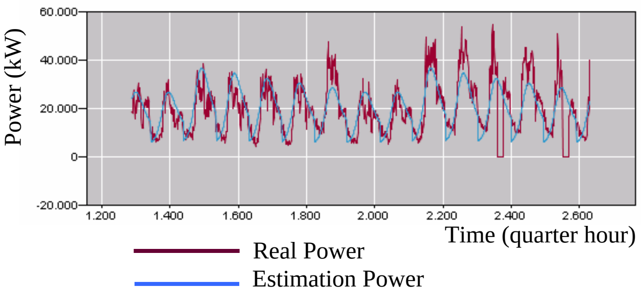


Figure 2- Estimation of a MV consumer consumption using a neural net.

By completing this missing data, the errors of the metered load curves are attenuated without making significant alterations in the real measures.

After the completion of the data, we prepared the latter 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 [15]. 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 are presenting the study performed for the period which consists of annual weekends.

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

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

Where (m) represents the consumer number in analysis, M represents the number of consumers of the sample and H = 96, represents the 15 minute intervals in a day.

The diagrams were computed using the field-measurements values, and, therefore, they need to be brought together to a similar scale for the purpose of their pattern comparison. This is achieved through normalization. For each consumer the vector l(m) was normalized to the [0-1] range by using the peak power of its representative load diagram [20]. This kind of normalization permits the maintenance of curve shape to compare the consumption patterns.

3.3 Data Mining Operations

3.3.1 Clustering

A clustering procedure based on a 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 [15] and [18]. 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 dataset. As the objective of our clustering is to find 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 established a number of 9 clusters. Besides, the partition precision will be proportional to the increase of the clusters number.

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 [12].

The clustering algorithm that produces the smaller MIA and CDI values prevails over the others in terms of performance. The smaller value of MIA indicates more compact clusters.

We obtained the clusters using the representative load diagrams and by dealing directly with the measurements power. The Two-step algorithm was applied to obtain the expected 9 clusters, for the two different load regimes.

With the resulting clusters achieved 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 representative load diagrams of the clients assigned to the same cluster.

Figure 3 shows the representative load diagram obtained for each cluster, using the measurement power directly on week days. For weekend days, we also obtained 9 representative load diagrams.

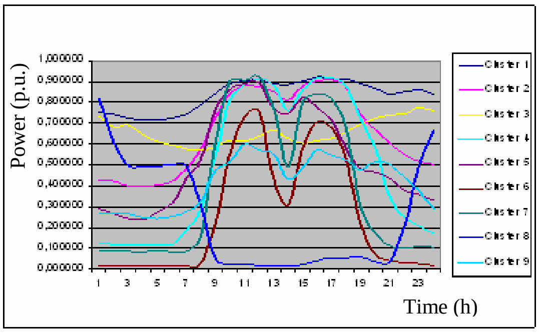


Figure 3- Representative Load Profile for each customer class for work days

As we can see in Figure 3, the clustering module is well separated from the customer population and the representative load diagrams were created with distinct load shape. Each curve that represents a cluster also represents a customer group with the same consumption pattern. A similar approach can be found in [16], [17] and [18].

In [16], 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 there is a poor correlation between the main clusters and the contractual data.

3.3.2 Classification

To obtain more relevant information to describe the consumption patterns of each population cluster, we have used a rule-based modeling technique, the C5.0 classification algorithm [20].

We have chosen this algorithm, in this first phase, due to it being easier to understand, since the rules derived from the model have a very straightforward interpretation.

The classification model should allow the attribution of a new consumer to a certain cluster, based on the rules generated by the classification model. Therefore, the rules must be intelligible and, for that to occur, normalized shape indicators were used as attributes in the classification model. The indexes vector f is formed by the indexes used in the studies accomplished in [18] and [20].

$$f = [f_1, f_2, f_3, f_4, f_5] \qquad (2)$$

Using the representative load diagrams, the load shape indexes are computed for each MV consumer. The indexes vector (2) that characterizes the representative load diagram shape was used in the classification model as an attribute in order to obtain intelligible rules.

Once again, the Two-Step algorithm was used in order to find the new 9 clusters for the week and weekend days, albeit now using the normalized shape indicators. These data sets were separate and formed a training group and a test group. The training group used by the classification model has 2/3 of the data, and the remaining 1/3 of the data was used for the test. The separation among test and training classes is to avoid the results being spoiled, so that the model error won't be an error influenced value. In [20], it was presented a rule set example obtained from the classification algorithm for the week data set. The obtained rules were simple and easy to understand.

The classification model used all the available attributes, for each rule selecting merely the attributes that provided larger information gain.

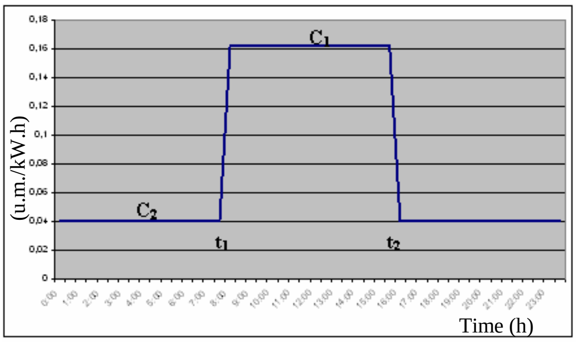
The model has been tested and its overall accuracy was approximately 95%, which shows that the results are satisfactory [20].

4 Development of New Tariff Structures

In [20] a new tariff structure was presented, taking into account the previous characterization of MV consumers on the basis of their electrical behavior. In that study, additional rates on power requested were ignored because these rates are the same for all customers and they are not related to the way that the customers consume electric energy.

The variation of the electric energy during a day in customer number 8 was calculated, belonging to cluster-2.

After that, the billing for that customer was calculated based on [12], using the regulated tariff. Following the shape of the representative load diagram of the cluster-2, which can be visualized in Figure 3, a new tariff structure was proposed with 2 different energy prices (C_1 and C_2), in such a way that the billing for the customer 8 will be the same when we apply the regulated tariff or the new tariff. The new energy price coefficients (C_1 and C_2) have been computed for the new tariff structures using the load profile.



m.u. – monetary units

Figure 4- Energy price variations for the new tariff structure.

This new tariff structure will be applied not only for the customer number 8 in study, but for all the customers that belong to the cluster-2.

5 Relationship between Customer and Electricity Supplier

The application of data mining techniques to the information that exists in the distribution companies will allow the identification of the consumers' classes. They can be identified as consumers with identical electricity consumption characteristics and also consumers with particular consumption characteristics.

This knowledge can be used by the retail companies essentially with 3 main objectives in mind:

1 - Identify the aspects that cause the increase of the diagrams peaks and develop strategies to lessen those peaks;

2 - Develop specific consumer’s contracts aspects, allowing for improvement of the electric nets used and the general quality of service.

3 – Allow the optimization of the offers of electric power purchase in the market and program the electricity production and maintenance actions more efficiently, in the case of having own electricity production.

In the electric power consumer's point of view, the typical profiles characterization can allow it to choose the electricity supplier with the best tariff structure proposal.

For the consumers, this knowledge also permits the modelling of their electric consumption habits, allowing a load elasticity, in other words, the customers can transfer their consumptions for schedules in which the electricity sale is cheaper.

They will also be able to celebrate contracts of electric energy interruption, sharing the risk of the prices volatility formed at the market with the electricity supplier.

6 Conclusion and further work

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

The results show 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. Normalized shape indexes were used as attributes in the classification model which automatically generated a rule set. These rules are simple and easy to understand.

By knowing the representative load diagram, and following the electrical behavior of the consumers, we presented a new tariff structure to apply for each customer class which must be sufficiently flexible to follow the variations in the load patterns of the customers.

Some advantages of the typical profile knowledge were presented and those ways can improve the relationship between consumers and the electric power supplier.

In terms of further work, our aim is to develop this work in the formulation of new tariffs structures in articulation with electricity markets.

ACKNOWLEDGMENT

The authors would like to express their gratitude to EDP Distribuição, the Portuguese Distribution Company, for supplying the data used in this work.

The authors would also like to acknowledge FCT, FEDER, POCTI, POSI, POCI and POSC for their support to R&D Projects and GECAD Unit.

References:

[1] Gellings, Clark W., "Emerging Energy Customers of the Twenty-First Century", CIGRE/IEEE Technical Session, *IEEE Power Engineering Review*, October, 1998.

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

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

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

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

[6] Baeza-Yates R. A., "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, 1992.

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

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

[9] Zhang, T., Ramakrishnon, R., and Livny M. 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, 1996.

[10]Chiu, T., Fang, D., Chen, J., Wang, Y., and Jeris, C, "A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment". Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining, 263, 2001.

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

[12]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, Nº1, February 2003.

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

[14]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.

[15]Rodrigues F, Duarte F.J, Figueiredo V., Vale Z., Cordeiro M.., "A Comparative Analysis of Clustering Algorithms Applied to Load Profiling.", *Proceedings of MLDM 03*, Leipzig, Germany, 2003.

[16]Figueiredo V., Rodrigues F., Vale Z. & Gouveia, B., "An Electric Energy Characterization Framework based on Data Mining Techniques". In the IEEE Transactions on Power Systems, Vol. 20, N.2, pp. 596-602, May 2005.

[17]Ramos, S., Figueiredo, V., Rodrigues, F., Pinheiro, R., Zita, V. "Characterization of MV Consumers Using Hierarchical Clustering", in Proc. of the International Conference on Knowledge Engineering and Decision Support, pp 199-204, Porto, Portugal, July, 2004.

[18]Sérgio Ramos, Fátima Rodrigues, Raul Pinheiro, Jorge Duarte & Zita Vale "*MV Electricity Contracts Definition Based on Patterns Extracted from MV Load Diagrams*", CEE´05 – 1st International Conference on Electrical Engineering, Coimbra, Portugal, 10 a 12 de October, 2005.

[19]Mcclanahan, R.H., "*Electric Deregulation*", IEEE Industry Applications Magazine, Vol. 8, nº2, pp. 11-18, March/April, 2002.

[20]Sérgio Ramos, Fátima Rodrigues, Raul Pinheiro, Judite Ferreira & Zita Vale "Decision Support System for Improving the Tariff Offer *Based on Patterns Extracted from MV Load Diagrams*", ICKEDS´06, in Proc. of the International Conference on Knowledge Engineering and Decision Support, pp 107-115, Lisbon, Portugal, May, 2006.