Low-voltage Electricity Customer Profiling based on Load Data Clustering*

Francesco Gullo, Giovanni Ponti, Andrea Tagarelli
Dept. DEIS, University of Calabria
87036, Arcavacata di Rende (CS), Italy
{fgullo, gponti, tagarelli}@deis.unical.it

Salvatore Iiritano, Massimiliano Ruffolo
Exeura S.r.l., Italy
{salvatore.iiritano, massimiliano.ruffolo}@exeura.com

Diego Labate
Enel Distribuzione S.p.a., Italy
diego.labate@enel.it

ABSTRACT

Current deregulated energy market requires that utilities have to face challenging issues that mainly arise from conceiving new customer-centric frameworks instead of early supplier-centric frameworks. Enel, a large international energy utility, is able to measure and store load profiles of their mass-market low-voltage (LV) customers in a flexible and effective way thanks to the well-established Telegestore project [3, 12].
In this paper, we present a study on the characterization of LV customers based on their consumption data. A time series based model is used to suitably represent load profiles and enable the detection of their characteristic trends. Besides this primary data, we also exploit meta-data associated to the load profiles, which is useful to enrich a-priori knowledge on the customers. We conceived a clustering framework for detecting groups of customers having similar consumption behavior. We experimentally evaluated the proposed framework on a real application concerning the characterization of Enel customers according to their load profiles. Preliminary experiments have shown results which are significant in terms of clustering validity and potentially useful to practitioners from the Enel utility.

1. INTRODUCTION

Today energy markets are characterized by a growing insecurity in the wake of their liberalization. Due to an increasing customer volatility, it is becoming more difficult for utilities to plan their investments through the next decades. In addition, the problem of characterizing and predicting their customers' behavior and fitting a proper tariff policy accordingly has been recognized as relevant in this context. Designing new tariff structures allows the energy utilities to encourage competition, efficiency, and economical use of the resources. Defining proper tariffs makes it possible to support customers' interests and, at the same time, to recover the cost of electricity in a reasonable time. Enel, a large international power utility, has recently completed an Italian project called the Telegestore project [3, 12]. By using up-to-date smart meters, Enel is able to measure and store load profiles of their mass-market LV customers in a flexible and effective way.

In recent years, technological improvement in electricity utility devices has leveraged various issues in load profile data management. In this respect, a significant research effort has been focused on load profile classification, especially regarding clustering of medium-voltage customers and short-term load forecasting of anomalous days. Customer classification puts the basis for properly designing tariff structures. The use of load pattern-based features has been identified as a key factor for classifying customers on the basis of their electrical consumption behavior. Classification allows utilities to promote collective tariffs rather than individual ones for each customer.

All the proposed techniques for load profile classification generally belong to pattern recognition and data mining approaches [6, 5, 4, 9, 1, 13, 14]. Load profiles are usually represented as time sequences and the notions of proximity used for comparing them are typically based on the Euclidean distance. In the context of load profile clustering, the most used approaches refer to partitional clustering and hierarchical clustering [8].

In this paper, we present a clustering framework for electricity customer load profiles, which is supported by information on meta-data (e.g., customer type, meter type, day, contract, location). Enel supported this work by providing data about 30,000 LV load profiles of anonymous Italian customers.

A major emphasis of our study was on the most typical class of electricity customers, i.e., private, residential domestic customers. Each customer load profile was segmented with respect to the type of day, which enabled a characterization of the customers' profiles on a per day basis.

We performed experiments by varying the algorithm and the distance measure in the proposed clustering framework. More precisely, we used the standard K-Means and the Euclidean distance as baseline method. However, we also resort to the Dynamic Time Warping distance, which is widely known to provide a better way to compare time series. Moreover, we introduce a simple top-down partitional algorithm, named TS-Part, which does not require the user to specify a desired number of output clusters, unlike the K-Means algorithm.

*Work supported by an ENEL-University grant under the project "EUREKA! An Idea for Energy – Profiling and Anomaly Detection in ENEL Customer Load Data"

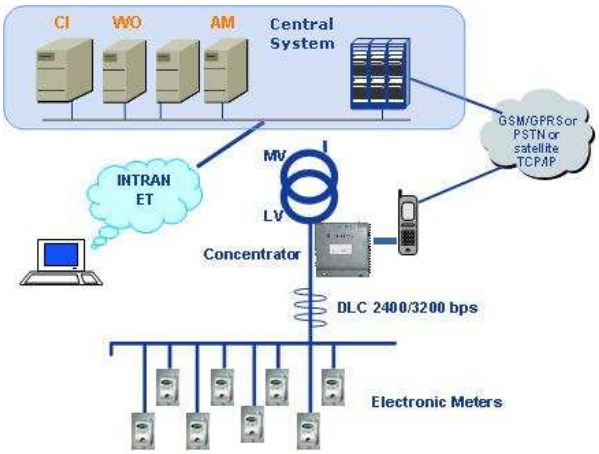

Figure 1: The Enel Telegestore architecture

Experimental results have shown that the Dynamic Time Warping supports higher-quality clustering than the Euclidean distance, in terms of both cluster separation and compactness. The best performance corresponded to setting TS-Part with the Dynamic Time Warping, which resulted in more clusters than those obtained by using the Euclidean distance. However, we observed that most of the data tend to group together in a relatively small number of clusters. This scenario enables the identification of relevant aspects which allow for supporting the design of tariff policies.

2. LOW-VOLTAGE ELECTRICITY CUSTOMER DATA

The Enel expertize in measuring and storing customer load profiles can be summarized in the Telegestore project [3, 12], whose conceptual architecture is shown in Figure 1. Communication between meters and concentrator is accomplished by a PLC (Power Line Carrier) channel, whereas the public GPRS/GSM Network is responsible for the communication between the concentrator and the central system. All energy related data are first collected from the smart meters by the concentrator. Then, such data are uploaded by the central system. The Enel Telegestore network devices consist of more than 31 millions of smart meters and more than 350,000 concentrators installed and remotely managed [12].

Enel smart meter is able to record and store active and reactive load profiles for all the four energy quadrants. A load profile represents the shape of the customer consumption chronologically ordered. Given a sampling period time, a smart meter logs the consumption corresponding to the associated location in a circular buffer. The sampling period is programmable and ranges from 1 to 60 minutes; as default, this is set to 15 minutes which allows for storing 38 days of load data, where each day is 96-sample long.

The smart meter also stores a flag register of "sample validity" in the stream of load data. This flag indicates a critical fault occurred during the sample measurement (e.g., a voltage interruption). In the experimental evaluation, we used this register to identify wrong samples and to correct each of them by using linear interpolation between the previous and the next valid sample.

3. TIME SERIES-BASED MODELING OF LOAD PROFILES

A time series T is a sequence $[(x_1, z_1), \ldots, (x_n, z_n)]$, where each pair (x_h, z_h) is comprised of a real numeric value (x_h) and a timestamp (z_h).

Similarity search and detection in time series databases relies on a measure of proximity among data points, which in principle should meet the following requirements: handling local time shifting, high efficiency in computation, low sensitivity to noise, and support for indexing.

In order to compare and measure the proximity between time series, the major approach consists in warping the time axis of one series to achieve the best alignment. The Dynamic Time Warping (DTW) algorithm has long been known in speech recognition [10], then was introduced to the data mining community as an effective solution to the sensitivity of the Euclidean distance to small distortions (i.e., fluctuations or phase shifts) in the time axis [2]. Given two sequences T_1 and T_2, DTW performs a non-linear mapping of one sequence to another by minimizing the total distance between them. For doing this, a $(|T_1| \times |T_2|)$-matrix storing the squared Euclidean distances between the two sequences is used to find an optimal warping path (i.e., a sequence of matrix elements) via a dynamic programming algorithm.

4. CLUSTERING LOAD PROFILE DATA

4.1 Algorithms

According to most of research works on clustering load profiles, we resort to the well-known paradigm of *centroid-based partitional clustering* [8]. Given a set of N data objects $\mathcal{D}$, the goal of a centroid-based partitional clustering is to partition $\mathcal{D}$ into a number $K < N$ of homogeneous subsets, called clusters, where each cluster is characterized by a data value, called centroid, which acts as a representative of that cluster. In this work we assume that the cluster centroids are computed as simple averages of the data (load profiles) in any specific cluster, since all the data have the same length in our setting. Of course, this assumption does not hold in general, and more refined methods for computing cluster centroids in time series data might be used [7].

The exemplary centroid-based partitional method is the popular K-Means algorithm [8]. In the experimental evaluation, we used the K-Means algorithm as baseline method. We also developed a top-down partitional algorithm, named *TS-Part*. A major feature of TS-Part is that the number of output clusters is not required as a parameter, rather it is determined during the clustering task. This represents an advantage in many real application contexts, like ours, in which there is no a priori information which guides the user to properly set the number of output clusters.

TS-Part starts by considering the input dataset as a single cluster, then two main steps are iteratively repeated until the convergence is reached. The first step consists in finding the best split for each cluster in the current clustering. The second step recomputes the cluster centroids and reassigns all data according to the current clustering, similarly to the K-Means algorithm. The convergence of the algorithm is reached when the split procedure does not perform any split.

In the splitting step, the quality of a given clustering solution is computed as the difference between the inter-cluster distance (i.e., the average pair-wise distance between all the cluster centroids) and the intra-cluster distance (i.e., the average distance between all the individual data within the cluster and the corresponding centroid).

The split operation hence depends on a threshold of minimum quality, which is initially set as the quality of the input clustering.

4.2 Assessment criteria

We evaluated compactness and separation of the solutions obtained by the clustering algorithms. More precisely, we employed two of the most used validity criteria in load profile clustering, namely Mean Index Adequacy (MIA) and Clustering Dispersion Indicator (CDI) (e.g., [5, 4, 13, 14]). Both criteria are based on information on the data to be clustered, the centroids of the clustering solution, and the number of desired clusters. MIA measures the compactness of a clustering solution by averaging the distances between each object within a cluster and its centroid. CDI expresses the degree of cluster separation as directly proportional to the average of the intra-cluster distance between the objects within the same cluster and inversely proportional to the pair-wise distances between the cluster centroids.

5. EXPERIMENTS

5.1 Data description and preparation

We were granted access to about $30,000$ Enel Italian LV customer load profiles, measured during the period between the first week of February 2009 and the last week of March 2009. All the load profiles have been provided in anonymous form.

The load profile set was preliminarily partitioned according to meta-data associated to each individual customer. Such meta-data represents commercial and technical extra attributes that Enel provided with each load profiles. Specifically, customer meta-data includes the following attributes:

- Meter type: specifies the power capacity and the number of phases (i.e, single-phase, multi-phase) of the meter associated to the customer;
- Contractual power: the maximum contractual power allowed to the customer;
- Contract date: the start date of the customer's contract;
- Commercial category: identifies the type of customer, including residential domestic, non-residential domestic, public lightning, etc.;
- Product category: identifies a particular (private or public) usage of the energy contract;
- Zone: refers to the geographical location of the customer.

According to the above information, we filtered in the available load profiles which correspond to the most common customer type, namely the "private", "residential domestic" customer. We considered only the active energy part of each load profile. The resulting $5,000$ load profiles were segmented in order to extract *daily* profiles. Since each daily profile is comprised of 96 samples, we obtained 30 daily profiles of 96 samples from each customer profile. Moreover, daily profiles were further partitioned depending on the type of day; precisely, we distinguished "weekdays" profiles from "saturdays" profiles and "sundays/holidays" profiles.

clustering algorithm	*distance measure*	*# of clusters*	*MIA*	*CDI*
K-Means	Euclidean	10	9.775	0.682
TS-Part	Euclidean	10	9.103	0.914
K-Means	DTW	66	7.986	0.008
TS-Part	DTW	72	**5.514**	**0.004**

Table 1: Best (average) performance of clustering: Weekdays load profiles

clustering algorithm	*distance measure*	*# of clusters*	*MIA*	*CDI*
K-means	Euclidean	19	7.456	0.100
TS-part	Euclidean	19	6.520	0.124
K-means	DTW	31	12.942	0.010
TS-part	DTW	37	**10.310**	**0.009**

Table 2: Best (average) performance of clustering: Saturdays load profiles

clustering algorithm	*distance measure*	*# of clusters*	*MIA*	*CDI*
K-means	Euclidean	17	9.964	0.109
TS-part	Euclidean	17	6.946	0.156
K-means	DTW	29	13.773	0.014
TS-part	DTW	32	**11.646**	**0.012**

Table 3: Best (average) performance of clustering: Sundays/holidays load profiles

5.1.1 The Rialto suite for data mining.

Experiments and analysis described in this work were conducted using **Exeura Rialto™** [11]. Rialto is a graphical environment for performing data mining and knowledge discovery tasks. In contrast to other similar data mining tools, Rialto contains most of the functionalities required by one user-friendly tool that allows users to design, create, explore, analyze, and execute data mining tasks, as well as to deploy predictive and descriptive models into other tools, applications, and systems. Thanks to the possibility of extending the capabilities of Rialto, it was possible to generate a set of ad-hoc plug-ins for managing the data from the Enel legacy repositories.

5.2 Preliminary results

We present here main results from clustering experiments on the three types of daily load profile sets, namely "weekdays", "saturdays", and "sundays/holidays". For each of the three cases, we performed multiple runs of both clustering algorithms (i.e., K-Means and TS-Part) and finally averaged the quality results, in terms of MIA and CDI, obtained over the runs. Each algorithm was equipped with Euclidean distance or DTW as distance measure. For each setting, the number of clusters was determined by TS-Part and then used to set the parameter (i.e., initial value of the number of output clusters) for the K-Means.

Tables 1–3 summarize the best (average) performance of the clustering algorithms obtained on the three cases. Using the DTW as distance measure mostly enabled either clustering algorithm to produce higher quality clustering solutions w.r.t. the ones obtained by using the Euclidean distance. This always holds in terms of CDI for all the cases, and also in terms of MIA for the "weekdays" case (which corresponds to the largest set of daily load profiles). The

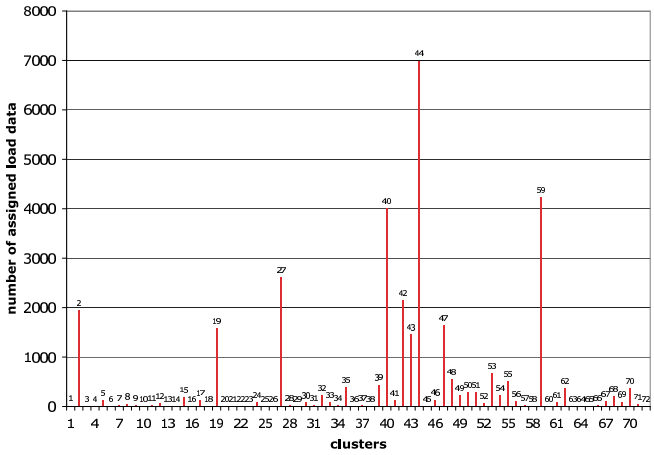

Figure 2: Distribution of weekdays load profiles over clusters obtained by TS-Part with DTW

better separation (CDI) obtained by using DTW reflects an increase in the number of clusters, which is explained by the fact that DTW is more sensitive than the Euclidean distance to time shifts and, consequently, it is capable of detecting more specific/descriptive clusters. The best results in each table correspond to the use of our TS-Part algorithm equipped with DTW. It should be noted that the better performance of TS-Part against K-Means concerns both MIA and CDI.

However, as we can see in Figure 2 referring to TS-Part with DTW, most of the profile data were assigned to a relatively small number of clusters, whereas a significant part of clusters likely corresponds to untypical habits of certain residential domestic customers (e.g., customers spending most of their time in residences which are located in places different from those declared in the contract).

It should be emphasized that the presence of a few, large clusters attracts major attention from a perspective of tariff policy design; conversely, the many, small clusters will be probably discarded. We indeed observed that, in typical behaviors of residential domestic customers, most of the energy consumption is concentrated on the lunch and dinner hours.

6. CONCLUSION

We presented a framework for clustering load profiles of electricity customers. Load data are managed and selected according to metadata associated to the customers, which mainly concern customer type, meter type, day, contract, and location. The proposed framework has been conceived to perform standard methods for clustering load profiles (i.e., K-Means and Euclidean distance). However, it also involves the well-known Dynamic Time Warping approach for comparing load profiles as time series. We also presented a new partitional algorithm which, unlike the K-Means algorithm, determines automatically the number of output clusters.

We performed experiments by differently combining clustering algorithms and distance measures. Experimental results have shown that the best performance is achieved by using the Dynamic Time Warping distance, which leads to better cluster separation and compactness. This also corresponds to clustering solutions having most of the data grouped together in a relatively small number of clusters. Such a scenario enables the identification of relevant aspects which allow for supporting the design of tariff policies.

We plan to extend the evaluation of our framework with more clustering methods and assessment criteria; in particular, we would like to exploit approaches specifically used in the power systems domain along with more general-purpose approaches from pattern recognition. More experiments are currently being carried out in order to consider both the active and reactive energy in the load profiles, and to enhance the evaluation of the framework by assessing its performances by significantly varying the sizes of the output clustering solutions.

7. REFERENCES

[1] D. Apetrei, I. Lungu, F. Batrinu, G. Chicco, R. Porumb, and P. Postolache. Load Pattern Classification and Profiling for A Large Supply Company. In *Proc. Int. Conf. on Electricity Distribution (CIRED)*, pages 1–4, 2007.

[2] D. J. Berndt and J. Clifford. Using Dynamic Time Warping To Find Patterns in Time Series. In *Proc. AAAI Workshop on Knowledge Discovery in Databases*, pages 359–370, 1994.

[3] B. Botte, V. Cannatelli, and S. Rogai. The Telegestore project in ENEL's metering system. In *Proc. Int. Conf. on Electricity Distribution (CIRED)*, 2005.

[4] G. Chicco, R. Napoli, and F. Piglione. Comparisons Among Clustering Techniques for Electricity Customer Classification. *IEEE Transactions on Power Systems*, 21(2):933–940, 2006.

[5] G. Chicco, R. Napoli, F. Piglione, P. Postolache, M. Scutariu, and C. Toader. Emergent electricity customer classification. *IEEE Proceedings Generation, Transmission & Distribution*, 152(2):164–172, 2005.

[6] V. Figueiredo, F. Rodrigues, Z. Vale, and J. B. Gouveia. An electric energy consumer characterization framework based on data mining techniques. *IEEE Transactions on Power Systems*, 20(2):596–602, 2005.

[7] F. Gullo, G. Ponti, A. Tagarelli, and S. Greco. A Time Series Representation Model for Accurate and Fast Similarity Detection. *Pattern Recognition*, In Press, available online at http://dx.doi.org/10.1016/j.patcog.2009.03.030, 2009.

[8] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, 1988.

[9] A. H. Nizar, Z. Y. Dong, M. Jalaluddin, and M. J. Raffles. Load Profiling Method in Detecting non-Technical Loss Activities in a Power Utility. In *Proc. IEEE Int. Power and Energy Conf. (PECon)*, pages 82–87, 2006.

[10] L. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Englewood Cliffs, N. J., 1993.

[11] Rialto. Exeura S.r.l., http://www.exeura.com/rialto, 2009 edition. Easy Analytics, Ready for Decision Making.

[12] S. Rogai. The Telegestore project: Progress and Results. In *Proc. IEEE Int. Symposium on Power Line Communications and Its Applications (ISPLC)*, 2007.

[13] G. J. Tsekouras, N. D. Hatziargyriou, and E. N. Dialynas. Two-Stage Pattern Recognition of Load Curves for Classification and Electricity Customers. *IEEE Transactions on Power Systems*, 22(3):1120–1128, 2007.

[14] G. J. Tsekouras, A. D. Salis, M. A. Tsaroucha, and I. S. Karanasiou. Load time-series classification based on pattern recognition methods. In Peng-Yeng Yin, editor, *Pattern Recognition Techniques, Technology and Applications*, pages 361–432. IN-TECH, 2008.