Classification of customers based on temporal load profile patterns

I. Benítez*, Instituto Tecnológico de la Energía, Spain
A. Quijano, Instituto Tecnológico de la Energía, Pain
I. Delgado, Instituto Tecnológico de la Energía, Spain
J.L. Díez, Universitat Politècnica de València, Spain

Summary

The deployment of Advanced Metering Infrastructure (AMI) is providing to utilities large amounts of energy consumption data from their customers, in form of daily load profiles with energy consumed per hour or a smaller period. These data can yield valuable results when analyzed, in order to extract useful knowledge about the typical patterns of consumption of energy from the customers. The proper mechanisms and tools have to be developed and implemented for this objective.

Big Data and Big Data Analytics systems will contribute to analyze this information and help to extract knowledge from the data, summarized in form of patterns or other mining knowledge, that will aid experts in decision support. In the present work a classification of customers based on their temporal load profiles is proposed. This classification procedure could be implemented in the current Big Data Analytics software systems, providing an added value to their statistical analysis options. Previous works in the literature present algorithms that allow to classify load profiles from customers by processing batch datasets and obtaining static patterns of load profiles. The proposed technique allows to analyze patterns not only in shape but also in their evolution or trend of energy consumption at each hour of the day through time. Specific quantitative indicators that characterize the patterns (and the consumers associated to them) are described and tested for this purpose.

1. Introduction

The current deployment of Information and Communication Technologies (ICT) to manage the electrical transmission and distribution grid, and the integration of the Advanced Metering Infrastructure (AMI) and Smart Metering from a growing number of customer premises, has brought to light the necessity of the development of adequate systems to gather, process and store huge amounts of data. For instance, one single smart meter can collect 96 quarter-hourly measures of active energy per day, 365 days per year, which yields an amount of 35040 values of energy per year, only for one client.

Utilities are solving this challenge by the implementation of Big Data systems and Big Data Analytics software tools, which are specifically designed to manage big amounts of data [1]. Big Data systems are built on new architecture paradigms such as cloud computing, and database systems such as NOSQL (Not Only Structured Query Language), which support the management of unstructured data. Big Data Analytics systems rely on the Big Data systems to gather, process and analyze the data, which can be obtained from distributed, scalable resources, or from the cloud, either in real time (i.e. "Stream" data) or off-line, even though stream mining systems are still in an initial phase of development [2]. Big Data Analytics systems, as indicated in [3], include three main functionalities:

- Data visualization, in form of charts, maps and new visualization designs that help to understand the results of the Big Data analysis.
- Statistical analysis, in form of summaries that describe collections of data, or inferential statistical analysis, that can be used to draw inferences about the process.
- Data mining, described as the computational process of discovering knowledge in large data sets. Usual algorithms for the processing of large data sets include

*ignacio.benitez@ite.es

KEYWORDS
Load profiles; Dynamic clustering; Pattern recognition; Classification.

classification, clustering, regression, statistical learning, association analysis and link mining.

The present work describes a classification system of customers based on their temporal load profiles, i.e., considering the daily load profiles through a sequence of consecutive days. In the data set analyzed, this period is of one year, but any other time frame could be considered (the previous week, for instance, or the past season). The development presented allows first to obtain patterns that summarize the energy consumption habits of a group of customers, and then allows to classify new customers to one of the previous patterns obtained by similarity in the energy consumption load profile. With this approach, the customers are classified not only by how they consume energy during the day, but also how they consume energy through a sequence of days. Each consumer is assigned to one of the resulting temporal patterns. The analysis allows the definition of quantitative indicators that characterize the patterns (and the consumers associated to them) from two perspectives:

- Load profiles (shape): indicators such as maximum daily consumption or peak/valley relation can be defined and computed.
- Trends (time): specific indicators are proposed to measure the trend of the pattern through time (e.g. increasing, decreasing, flat), and the amount of variations between sequential days, as some sort of dynamics measure.

Obtaining these indicators allows to globally describe the consumption of a number of consumers in a given period of time at a glance: the visualization of the patterns, such as the example of Figure 1, will allow an experienced user to clearly identify the common patterns of energy consumption, with the higher number of consumers assigned to them, and to quickly identify unusual patterns with unexpected behaviours or with fewer number of consumers assigned to them, and to think about the possible reasons behind this behaviour. Besides, the computation of the previously commented indicators will allow a quantitative characterization of the patterns, which can be used as an input for a classification system. This system will classify and identify consumers given some specific criteria for the classification, which must be defined in form of rules or threshold values for the indicators and the number of consumers at each pattern.

2. Segmentation of customers based on temporal load profiles

A dynamic clustering technique is applied on the data set of customers' load profiles. This algorithm is called END-KMH (Equal N-Dimensional K-means Hausdorff-based), and has been designed for the clustering of time series data with *n* dimensions or characteristics, such as load profiles time series (24 in this case).

The energy consumption profiles from the customers are seen as 3D shapes, defined by the 24 hours load profile, where each hour is considered as a feature or dimension of the data, with a length equal to the number of days considered in the analysis (one year in the present work). The time series clustering algorithm needs to define a measure of similarity between each client's profile and all the resulting centroids at each iteration. The Euclidean distance is usually used for this in most works. The proposed solution in this work is a two-step process. First, all the shapes are decomposed in a number of linear surfaces, by applying least squares regression. The number of surfaces and the vertices is predefined, based on the expert's knowledge of the typical behavior from residential users regarding energy consumption. Then, the resulting surfaces are compared by computing the Hausdorff distance [4] between them, and a global similarity value is obtained, given by the average value of all the Hausdorff distances between the different surfaces. A more detailed description of this algorithm can be found in [5], where the same algorithm has been applied with the objective to identify uncommon patterns for reliability and maintenance purposes.

For the present work, a data set has been selected to serve as an example of the capabilities of the analysis proposed. This data set is the "SmartMeter Energy Consumption Data in London Households", provided by the Greater London Authority through the London Data Store web site. This data set gathers half-hour energy consumption measures from a sample of 5567 customers between years 2012 and 2014. For this analysis, the hourly energy consumption values have been processed from the first 2000 households through year 2013 (therefore 365 days x 2000 households =

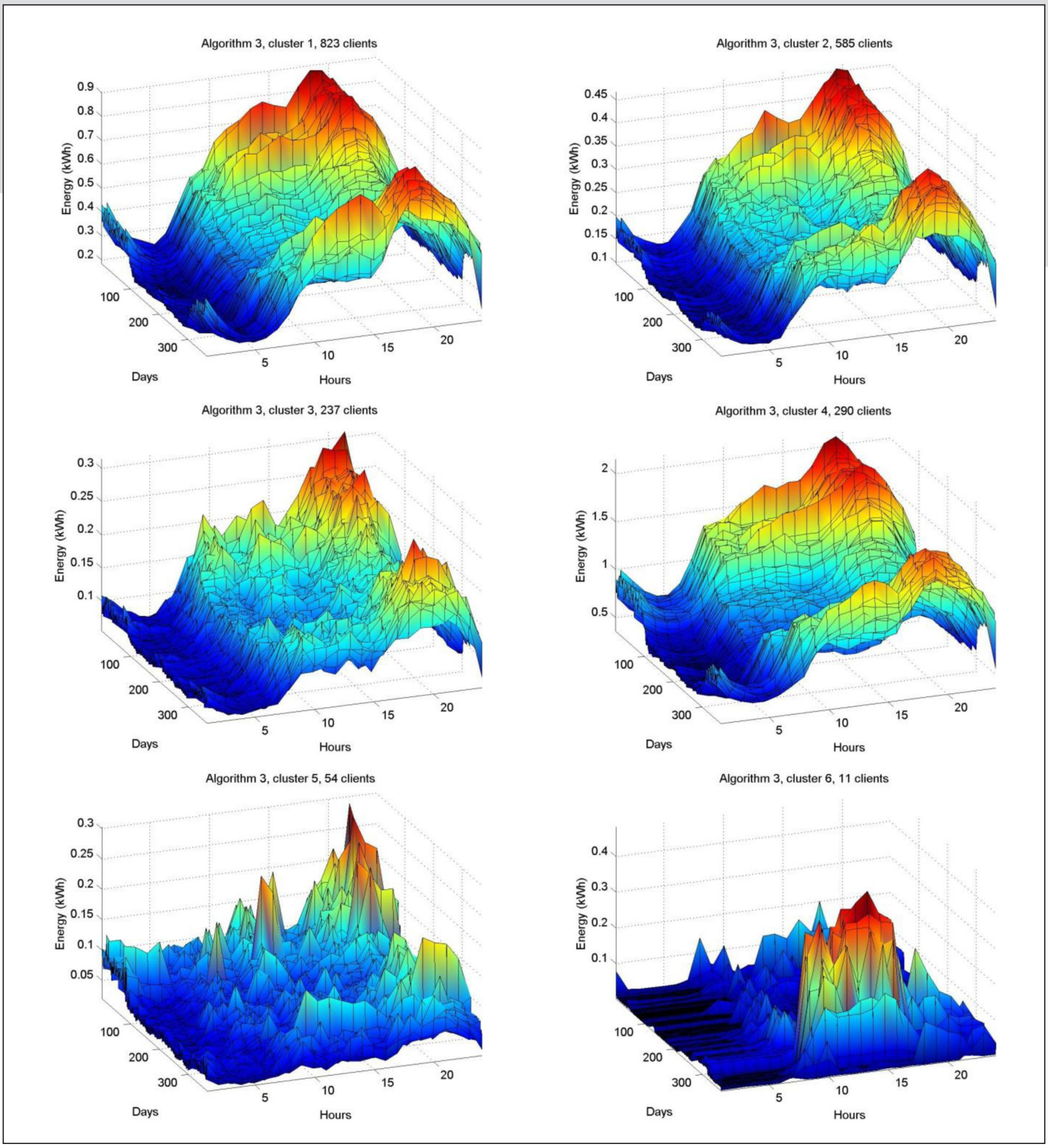


Figure 1. Example of six temporal clusters (and patterns) obtained from the first 2000 households of the Smart Meter Energy Consumption dataset for year 2013, provided by the Greater London Authority.

730000 load profiles being analyzed). Figure 1 depicts the patterns obtained after applying the dynamic clustering algorithm previously described. The number of clusters to be found is a parameter that must be defined. A number of 6 clusters to be found has been chosen, based on previous experiences with the data.

Observing the rasulting patterns in Figure 1, some quick conclusions can be derived. It can be observed, for instance, that there are four clusters which gather the majority of households (clusters 1 to 4), and there are two clusters with probably unusual energy consumption patterns and a more reduced number of customers or households belonging to them. These are clusters 5 and 6, with 54 and 11 customers respectively. Moreover, observing clusters 1 to 4, another conclusion could also be obtained: that the shapes of the four clusters are very similar, and their main difference is in the average energy consumption per day, ranging from a maximum peak of around 0.3 kWh in households from cluster 3, to a maximum peak of energy consumption of around 2 kWh in households from cluster 4. The seasonality effect can also be observed in the six clusters obtained, with an elevated rate of energy consumption in Winter (around January and February months).

Time series clustering allows, therefore, obtaining patterns that evolve through time, in a time frame defined by the expert. This allows an interpretation of the results that depicts the full dynamic behavior of all the objects, allowing a much more complete (and also complex) global view of the resulting temporal patterns.

Following, the resulting patterns are characterized by means of specific quantitative indices that allow to describe the resulting patterns in their evolution in shape and through time.

3. Definition of indices to characterize temporal load profiles

In order to characterize the resulting patterns of temporal load profiles, the following qauntitative indices have been described:

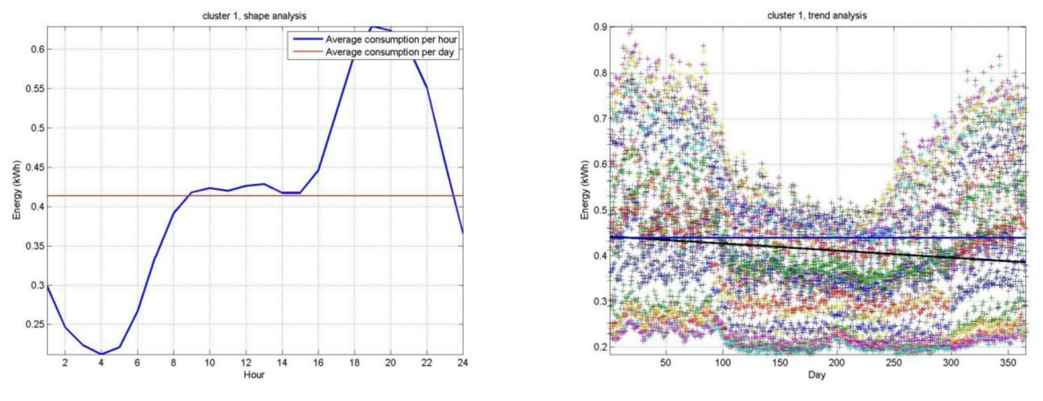
- Shape analysis: daily average energy consumption. This index is computed as the average value of all the average values of energy consumption at each hour of the day through the year. These are the values that can be seen depicted in the example of Figure 2 (a).
- Shape analysis: Relation between peak and valley energy consumption. This index is computed as the difference between the maximum value of hourly energy consumption from the whole year data and the minimum value of hourly energy consumption.
- Daily evolution: consumption trend through the year. A simple linear regression model is obtained, by applying the least squares algorithm, to fit the data of energy consumption, arranged as sequences of daily values at the same hour of the day, to a lineal model with two coefficients, being the slope and an independent variable. The regression model is depicted in black color in the example of Figure 2 (b).

Daily evolution: RMS through the year. This index gives information regarding the amplitude of the daily energy consumption values through the year. It is computed as the Root Mean Sqaure (RMS) of all the energy consumption values. The result can be observed in blue colour, in Figure 2 (b).

Other indices could be defined, for instance a new index to describe the seasonality variation. This will be a subject for further developments. Table I displays the resulting indices from the patterns previously obtained. As can be seen, based on the analysis of the resulting patterns, the six clusters, and the customers that belong to them, are characterized quantitatively in terms of how they consume the energy during the day, and also their hour by hour dynamic evolution through the year.

Cluster	No. of customers	Shape		Daily evolution	
		Daily average (kWh)	Peak – valley (kWh)	Trend through year (slope)	RMS through year (kWh)
1	823	0.414	0.718	-0.00015	0.439
2	585	0.214	0.369	-0.00007	0.226
3	237	0.123	0.258	-0.00006	0.130
4	290	0.897	1.789	-0.00057	0.960
5	54	0.064	0.283	-0.00009	0.069
6	11	0.022	0.480	0.00014	0.049

Table I. Clusters and quantitative indices computed from the patterns.



(a)

(b)

Figure 2. Graphical representation of the quantitative indices defined to describe the temporal patterns obtained, in terms of shape analysis (a) and the evolution of the load profile through the days (b). Cluster 1.

The next step of the analysis is to classify new customers on the clusters previously obtained, based on similarity to one of the six patterns detected. This step is described next.

4. Classification of new customers

Whenever time series load profiles from new customers are available, these can be compared with the six patterns previously obtained, and the new customer can be assigned to the most similar one, based on a measure of mathematical similarity. This classification will have a certain degree of error, that can be evaluated by computing the same quantitative indices for the custome's load profile, and comparing the values with the assigned pattern ones. Global error indices, such as the MAE (Mean Absolute Error) can be computed for each pattern based on the classification of new customers.

As a validation example, a test has been developed with the six patterns obtained from the first 2000 households of the Greater London Authority data set. A set of new 1000 households from the same database has been classified to one of the six patterns. A modified Euclidean distance has been used to compare each customer's time series load profile with each of the six patterns. The computation of this similarity measure is shown in Equation (1), where *n* is the number of features or characteristics of the data (24 hours in this case), and B is the norm. Since the Euclidean distance is used, the identity matrix has been applied as the norm.

$$d(X_i, V_J) = \frac{1}{n}\sum_{k=1}^{n} \|X_{ik} - V_{jk}\|_B^2 = \frac{1}{n}\sum_{k=1}^{n}\left((X_{ik} - V_{jk})^T B(X_{ik} - V_{jk})\right) \qquad (1)$$

The 1000 new households are classified to the pattern with the highest similarity, as computed by the distance in Equation (1). Following, the quantitative indices are computed for all the customers time series load profiles, and a global error measure in the classification is computed for each pattern, as the mean of the difference between the indices from the customers assigned to the pattern and the pattern itself. This error is computed as absolute (i.e., the MAE) in all the indices except the slope or trend. Table II shows the results obtained.

Comparing these results with the values of the indices in Table I it can be observed that the errors obtained in the classification are low in most of the indices, for all the six clusters. The only exception could be the peak – valley relation, which is high in all the cases, probably due to the disparity in energy consumption time series profiles from the customers, compared to the averaged patterns of consumption.

5. Conclusions

The classification system presented can automatically process the temporal patterns from a given time frame, and summarize a big number of load profiles from customers in summarized groups of consumers with a specific behaviour in shape and time. Applications of this system can be diverse: since the selection of customers for demand side management purposes, to the design of dynamic or temporal maps of energy consumption that take into account the computed trends in increasing or decreasing energy consumption.

The definition of quantitative indices computed from the resulting patterns allow to characterize these patterns, and the customers associated to them, in the shape of their load profiles and the variation of these load profiles

Cluster	No. of new customers classified	Shape		Daily evolution	
		MAE Daily average	MAE Peak – valley	Error in trend through year	MAE RMS through year
1	399	0.099	2.996	-2.29 x 10^{-05}	0.205
2	309	0.055	2.221	2.88 x 10^{-05}	0.135
3	125	0.030	1.722	2.26 x 10^{-05}	0.094
4	132	0.210	4.038	9.79 x 10^{-05}	0.292
5	32	0.026	1.413	-2.64 x 10^{-05}	0.084
6	3	0.021	0.333	1.52 x 10^{-04}	0.043

Table II. Results of classification of new customers and classification errors per index.

through the days. From the temporal patterns obtained, new customers can be classified to them, according to mathematical similarity. The classification errors can be computed from the defined quantitative indices, allowing to validate the process.

6. Bibliography

[1] *C. P. Chen and C.Y. Zhang "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data" (Information Sciences , 2014, vol. 275, pages 314-347)*

[2] *K. Kambatla, G. Kollias, V. Kumar and A. Grama "Trends in Big Data analytics" (Journal of Parallel and Distributed Computing , 2014, vol. 74, pages 2561-2573)*

[3] *H. Hu, Y. Wen, T.S. Chua and X. Li "Toward Scalable Systems for Big Data Analytics: A Technology Tutorial" (Access, IEEE, 2014, vol. 2, pages 652-687)*

[4] *F. Hausdorff "Grundzüge der Mengenlehre" (Veit and Company, 1914)*

[5] *I. Benítez, A. Quijano, J.L. Díez and I. Delgado "Clustering of Time Series Load Profiles for Grid Reliability" (CIGRE International Conference on Condition Monitoring, Diagnosis and Maintenance - CMDM 2015)*

7. Biographies

Ignacio Benítez

PhD in Automation and Industrial Computer Science from the Universitat Politècnica de València in 2016. Automation and Industrial Electronics engineer from the Universitat Politècnica de València. In October 2008 he joined the Instituto Tecnológico de la Energía as R+D+i technician. He is currently part of the Smart Grids department, where he develops projects related to sustainable mobility, analyzing the impact of the integration of the EV in the low and medium voltage grid. Previously, he was responsible for the Coordination and Development of actions undertaken in the area of Advanced Control Techniques, addressed to the application of control methodologies and knowledge aimed at achieving goals related to energy efficiency and integration of renewable energy resources. He has proven experience in robotics, artificial intelligence, machine learning, control systems engineering, fuzzy logic, prediction models, neural networks, data mining and pattern recognition techniques (clustering).

Alfredo Quijano

Alfredo Quijano was born in 1960 in Valencia, Spain. He received the Electrical Engineer degree and the Ph.D. degree from Universitat Politècnica de València, in 1986 and 1992, respectively. He is the CEO of the Instituto Tecnológico de la Energía – Universitat Politècnica de València and Vicepresident of ITE. He is also a teacher and researcher at the Universitat Politècnica de València in the Electrical Engineering Department. His current research activity is focused on applied research for the Energy area and the electrical technology including renewable energies, high voltage, metrology, new materials and applications and research results transfer to companies.

Ignacio Delgado

Mr. Ignacio Delgado is Industrial Engineer with over six years of industry experience. From May 2006 to August 2016, he was part of the ITE team, developing his work as head of the department of Smart Grids of ITE, working at management and development of national and European projects on smart grids, Renewable Energy Integration network, Distributed Generation, Demand Management and Automation and Remote management of the network.

In recent years his research has focused on projects on the active management of demand in the domestic sector, analyzing and studying the functional requirements to integrate systems on networks to provide all of greater intelligence to manage and behavior as the real-time user consumption. He has recently moved to Glasgow, to work in the Centre of Excellence of Scottish Power Energy Retail.

José Luis Díez

José Luis Díez received a M.Sc degree in Industrial Engineering in 1995, and the Ph.D. degree in Control Engineering in 2003, both from the Universitat Politècnica de Valencia, Spain. He is currently an Associate Professor and he has been teaching since 1995 at the Systems Engineering and Control Department, Universitat Politècnica de Valencia, in a wide range of subjects in the area such as automation, linear systems control theory, digital signal processing, and intelligent systems. His research interests include complex systems modelling and identification (biomedical, biological, energy and social systems), data mining, clustering techniques, intelligent control, and control education.