Electric Power Systems Research 140 (2016) 517–526

Contents lists available at ScienceDirect

Electric Power Systems Research

journal homepage: www.elsevier.com/locate/epsr

Dynamic clustering of residential electricity consumption time series data based on Hausdorff distance

Ignacio Benítez [a,*], José-Luis Díez [b], Alfredo Quijano [a], Ignacio Delgado [a]

[a] Instituto Tecnológico de la Energía, Avda. Juan de la Cierva 24, 46980 Paterna, Spain
[b] Department of Systems Engineering and Control, Universitat Politècnica de València, Camino de Vera 14, 46022 Valencia, Spain

A R T I C L E I N F O

Article history:
Received 26 October 2015
Received in revised form 25 April 2016
Accepted 20 May 2016
Available online 3 June 2016

Keywords:
Dynamic clustering
Data mining
Load profiles

A B S T R A C T

As the analysis of electrical loads is reaching data measured from low voltage power distribution networks, there is a need for the main agents involved in the operation and management of the power grids to segment the end users as a function of their shapes of daily energy consumption or load profiles, and to obtain patterns that allow to classify the users in groups based on how they consume the energy.

However, this analysis is usually limited to the analysis of single days. Since the smart metering data are time series formed by sequential measurements of energy through each hour or quarter of hour of the day, and also through each day, thanks to the implementation of Advanced Metering Infrastructure (AMI) and the Smart Grid technologies, it becomes clear that the analysis of the data needs to be extended to consider the dynamic evolution of the consumption patterns through days, weeks, months, seasons, and even years.

This is the objective of the present work. A new framework is presented that addresses the dynamic clustering, visualization and identification of temporal patterns in load profiles time series, fulfilling the detected gap in this area. The present development is a generic framework that allows the clustering and visualization of load profiles time series applying different classical clustering algorithms. A novel dynamic clustering algorithm is also presented, based on an initial segmentation of the energy consumption time series data in smaller surfaces, and the computation of a similarity measure among them applying the Hausdorff distance. Following, these developments are presented and tested on two dataset of energy consumption load profiles from a sample of residential users in Spain and London.

1. Introduction

The European Technology Platform on Smart Grids (ETP SG) has issued a report in 2015 [1] on the research and development needs foreseen by the platform for the EC Horizon 2020 Research and Innovation Programme [2], for the years 2016 and 2017. One of the main challenges identified by the ETP SG is the utilization of smart metering data. According to the ETP SG, “a very large amount of data is being collected whose potential has been untapped”.

The question arises on how the large amounts of smart metering data can be used in a way to be profitable to an interested party or agent. Data mining techniques can provide the tools to achieve this objective. The term “data mining” gathers a number of different algorithms and techniques which have as objective the analysis and extraction of useful information from large sets of data [3].

Han and Kamber [3] define two main objectives of the data mining process, as a function of the data mined and the kind of knowledge sought. These objectives are the static analysis, or the analysis of static data, and the evolution analysis, where the trend of the series and the temporal evolution of the data is a key factor in the objective of the analysis.

The development presented in this paper allows to obtain 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, therefore providing a much more complete (and also complex) information, from where conclusions can be obtained and actions can be determined, to fulfill the purposes of the data mining process. Issues such as identifying specific groups with special trends or shapes in time, or comparing clusters’ differences according to their entire behavior in the time frame, can now be considered.

First, a state of the art is presented, regarding dynamic clustering algorithms found in the literature and previous clustering analyses regarding the segmentation of load profiles. Following, the

* Corresponding author. Tel.: +34 961366670; fax: +34 961366680.
E-mail address: ignacio.benitez@ite.es (I. Benítez).

http://dx.doi.org/10.1016/j.epsr.2016.05.023
0378-7796/

Nomenclature

$\bar{X_{i_k}}, \bar{V_{j_k}}$	average values of the X_{i_k} and V_{j_k} time series
μ_{ij}	membership value of object i to cluster j
ε_i	residual or model error at time or instant i
B	norm matrix
c	number of clusters or classes
$d(X_i, V_j)$	distance or similarity function between X_i and V_j
m	degree of fuzziness of the clusters (usually a value higher than 1)
n	number of features or characteristics of the data
p	total number of time samples or instants
V_j, V_k	centroids or prototypes of the classes or clusters j and k
X_i	feature vector of object i
$W=[w_0\ w_1\ w_2]^t$	vector of coefficients of a linear surface model

development made is presented and tested on two different datasets of load profiles from a sample of residential low voltage consumers, in Spain and London. The results are described and discussed. Finally a conclusions section is included.

2. State of the art on dynamic clustering techniques applied on load profiles

With respect to the dynamic nature of the data and the cluster analysis, Weber [4] classifies the cluster analysis in four types or categories, according to the dynamic nature of the data and the clusters:

- Type 1. The data is treated as static and the clustering process is also static.
- Type 2. The data is treated as static but the number of clusters may vary at each new computation. In this system, issues such as clusters formation, collapse, split or fusion must be considered.
- Type 3. The data is treated as dynamic, evolving through time, as trajectories of the different data features or dimensions through time. The number of clusters is fixed. The resulting centroids or patterns are therefore defined by feature trajectories that evolve through time. The present work approaches this type of dynamic clustering analysis.
- Type 4. The data is also treated as dynamic, as in Type 3, becoming feature trajectories that evolve through time, and the number of clusters varies dynamically at each iteration. Clusters and patterns can, therefore, as in Type 2, merge or split.

Liao [5] makes a differentiation of clustering types for time series data based on three main approaches: clustering on the raw data, clustering on a feature-based transformation of the data, and clustering on a model-based transformation of the data. The clustering algorithms presented in this paper are based on the analysis of the raw time series data, therefore a brief review on the current state of the art in this field is described next.

A number of time series clustering algorithms are based on modifications of the K-means [6] or Fuzzy c-means or FCM [7] algorithms. Weber [4] describes the algorithm called Functional Fuzzy C-means or FFCM, as a time series generalization of the FCM. The FFCM algorithm presents a modified calculation of the membership value at each iteration, indicated in Eq. (1), where the distance function d is based on fuzzy inference.

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d(X_i, V_j)}{d(X_i, V_k)} \right)^{2/(m-1)}} \tag{1}$$

Regarding recent years, Izakian et al. [8] present a clustering algorithm for spatiotemporal data where the Euclidean distance is replaced by an "augmented" distance, which is the weighted sum of two Euclidean distances: the comparison of the spatial components and the comparison of the temporal features. This modification is later extended by Izakian and Pedrycz [9] to n features or dimensions with the concept of blocks or groups of similar features, computing a weighted sum of Euclidean distances where the different weights are obtained by Particle Swarm Optimization (PSO) [10].

Concerning the analysis of load curves of energy consumption, all the works found in the literature correspond to static or Type 1 clustering. Realizing that no specific development had been found addressing the dynamic clustering and visualization of energy consumption load profiles time series data, the authors of the present work presented a previous paper [11] where a dynamic K-means clustering algorithm was developed, by modifying the static K-means algorithm to obtain the similarity distances among objects taking into account all the Euclidean distances between each pair of objects from their coincident time stamps.

3. Development of algorithms and techniques to perform dynamic clustering on load profiles time series data

The new approach presented in this work applies the concept of Type 3 dynamic clustering described. In this case the feature or dimension *trajectories* of the objects are clustered. The dimensions of two objects are compared as sequences of n samples, and a final distance is obtained as the average of all the comparisons of features at the 24 dimensions. Although the two approaches may deal similar mathematical results, they are quite different and, depending on the operators and similarity measures used, may produce very different outcomes. The first approach can be seen as a succession of Type 1 static clustering calculated for n times and clustered together. The second approach is designed as a Type 3 dynamic clustering of dynamic trajectories through time, with a fixed number of classes.

The similarity measure proposed can be seen as an "augmented" distance, in the sense described by Izakian et al. [8] or Izakian and Pedrycz [9], since the similarity in static or Type 1 clustering is augmented to process time instants in n features or dimensions. It can also be seen as a development based on the description of the membership function of a time series object to a class made by Weber in the FFCM algorithm [4]. The distance function operator d is replaced by specific distance functions able to compare two time series and yield a value of similarity. None of the previous works presented, however, have been developed to analyze and visualize the resulting clusters in the form of n or, in this case, 24 dimensions of dynamic data objects.

Moreover, a new dynamic clustering procedure is presented, as a modification of the static K-means algorithm, but applying an initial decomposition of the data object in smaller linear surfaces and comparing them applying a Hausdorff-based similarity distance. These developments are presented next.

3.1. Development of cluster validity indices for the evaluation of dynamic clustering on time series n-dimensional data

A number of cluster validity indices are used for the comparison of the results of dynamic clustering algorithms on time series