Energy and Buildings 91 (2015) 16–25

Contents lists available at ScienceDirect

Energy and Buildings

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

A data-mining-based methodology to support MV electricity customers' characterization

Sérgio Ramos, João M. Duarte, F. Jorge Duarte, Zita Vale*

GECAD, Knowledge Engineering and Decision Support Research Center, Polytechnic of Porto (ISEP/IPP), R. Dr. Antonio Bernardino de Almeida, 431, 4200-072 Porto, Portugal

ARTICLE INFO

Article history:
Received 17 September 2014
Received in revised form 19 January 2015
Accepted 20 January 2015
Available online 29 January 2015

Keywords:
Load profiling
Data mining
Clustering
Classification
Clustering validity

ABSTRACT

This paper presents an electricity medium voltage (MV) customer characterization framework supported by knowledge discovery in database (KDD). The main idea is to identify typical load profiles (TLP) of MV consumers and to develop a rule set for the automatic classification of new consumers. To achieve our goal a methodology is proposed consisting of several steps: data pre-processing; application of several clustering algorithms to segment the daily load profiles; selection of the best partition, corresponding to the best consumers' segmentation, based on the assessments of several clustering validity indices; and finally, a classification model is built based on the resulting clusters. To validate the proposed framework, a case study which includes a real database of MV consumers is performed.

1. Introduction

Concerning the electricity market environment, the characterization of the electrical consumers assumes an important supporting tool for electric utilities, for understanding and predicting the behaviour of their electricity customers. It is expected for suppliers to know, as much as possible, the electrical consumption habits of their customers, to offer them suitable electric energy services at the least cost and thus differentiating themselves from others competitors. The knowledge about customers' consumption patterns is particularly important for setting up dedicated commercial offers. Indeed, load patterns are broadly used in tariff design, system planning, system maintenance, load management and marketing [1]. Typically, the electrical suppliers companies cluster consumers into representative classes and use the representative load profile to study consumers' behaviour [2–5].

Automatic meter reading (AMR) systems, normally operating at quarter-hour intervals, have been implemented by most of electric power companies [6], mainly for MV customers. In fact, the European Union's current strategy promotes its utilization [7]. So, gradually, a huge amount of data concerning electricity consumption will become available and stored into databases, allowing load patterns to be extracted from these. In the deregulated electricity industry there is a distinct separation throughout the value chain of the power system: production, transmission, distribution and retail. While transmission and distribution companies explore the distribution power network, according different voltage levels, the retail companies are responsible for managing relations with end consumers, including invoicing, billing and customer services, and have some flexibility in formulating the tariff offers, assuring that their offers meet the requirements by the regulatory authorities in the form of prices [5,8].

Conceptually, the tariffs offers are formulated with reference to a specific consumer's class, defined by a set of technical and commercial attributes. The distinction between customers' groups can be made based on the definition of macro-categories, e.g., residential, commercial, industries, public lighting, or others specific consumers. There are also other attributes that can be used for the distinction of customers' classes, such as the contracted power value, annual energy consumption and the voltage level or, as presented in [9], a criterion based on the cost of energy purchased from the pool market by a retailer. However, in large number of research works in this field of study, the load profile for tariffs purpose is typically performed with load data. Also, the customers' characterization could be accomplished, for example, based on the commercial type of activity. However, the load profiles that belong to the same commercial type of activity reveal different electrical consumption habits. Thus, using the commercial type of activity for customers' categorization is generally not efficient for representing

* Corresponding author. Tel.: +351 22 8340500; fax: +351 22 8321159.
http://www.gecad.isep.ipp.pt/.
E-mail address: email.zitavale@gmail.com (Z. Vale).

http://dx.doi.org/10.1016/j.enbuild.2015.01.035
0378-7788/

the electricity consumption [4,5,8]. For the electricity customers without measured data available, their association with one of the formed typical load profile classes can be identified *à posteriori* based on available information and attributes of that customer and of the obtained customer classes. Power utilities can also obtain load profiles from AMR customers and the so-called virtual load profile (VLP) from non-AMR customers in order to create load profile of all customers [5,6].

In the last years, dedicated research effort has been developed in order to study load profiling. Typically, pattern recognition methods have been applied to electricity consumption data. A variety of clustering algorithms have been proposed to group together load diagrams with similar shapes. In [10] it is possible to find a brief overview of well-known clustering methods, discussing its major challenges and some of the emerging and useful research directions are pointed out.

This paper presents a data-mining-based methodology to identify typical load profiles, using a real database provided by the Portuguese utility. To conduct data partitioning, several clustering algorithms have been used and the evaluation of the quality of the obtained data partitions were assessed by cluster validity indices. The implemented methodology is extremely useful for electrical suppliers' companies, as well as consumers' aggregators, to identify the typical daily load profile supporting the design of new tariff structures and to improve their strategy of market share, either by optimizing their power purchase option, or by the definition of demand response programs. A new customer can easily be placed in one of the defined clusters using a classification model. With a significant increase of clients, one may need to start the whole clustering process to find the new optimal data partition, which can be seen as a limitation of the proposed approach.

The remaining of this paper is organized as follows. In Section 2 a review of data mining techniques is presented. Section 3 addresses the proposed methodology for electrical customers' characterization and classification. In Section 4 a case study using real data is presented. The last section summarizes the concluding remarks.

2. Data mining techniques

Data mining is the task of discovering patterns in large data sets involving methods of artificial intelligence, machine learning, statistics, and database systems. In this section, a brief description of some methods used for data clustering analysis and classification is presented.

2.1. Data clustering algorithms

Clustering is the process of partitioning a set of data objects into clusters based on a concept of similarity or proximity among data. Even though there is a huge number of clustering algorithms in the literature [11,12], no single algorithm can effectively find by itself all types of cluster shapes and structures.

The purpose of any clustering technique consists in dividing a data set *Xcomposed* of n data patterns $\{x_1, \ldots, x_n\}$ into K clusters $\{C_1, \ldots, C_K\}$, such that similar data patterns are placed in the same cluster and dissimilar data patterns are grouped into different clusters. The set of clusters $P = \{C_1, \ldots, C_K\}$ is referred as data partition. The major clustering algorithms can be classified into the following categories:

I. Partitive algorithms initially define K seed points $\bar{x}_k$ (centroids or medoids), one for each cluster, and iteratively update these points to optimize some objective function. At each iteration, each object x_i is assigned to the most similar seed point. Three **partitive** algorithms are shortly described ahead.

The K-Means algorithm [13] is the best known data clustering algorithm. K-Means tries to minimize the within-cluster sum of squares $\left(\left(\sum_{k=1}^{k}\sum_{x_i \in C_k}\left\|x_i - \bar{x}_k\right\|^2\right)\right)$ where $\left\|x_i - \bar{x}_k\right\|^2$ is the Euclidean distance between pattern x_i and its closest cluster centroid $\bar{x}_k$.

This algorithm takes as parameter the desired number of clusters K and randomly chooses K data patterns as the initial centroids $\{\bar{x}_1, \ldots, \bar{x}_K\}$ of each cluster. Then, K-Means algorithm iterates between two steps: find for each pattern $x_i \in X$ the closest centroid $\bar{x}_k$ and assign it to the corresponding cluster C_k, and update each centroid $\bar{x}_k$ as the mean vector of the corresponding cluster. This process is repeated until no pattern assignments are changed from one iteration to the next one, meaning the algorithm converged to a (local) minimum.

Clustering process can be made given some constraints between data patterns (pairwise constrained clustering). The PC-KMeans algorithm (PCKM) [14] formulates the goal of clustering in the pairwise constrained clustering framework as minimizing a combined objective function, which is defined as the sum of the total square distances between the points and their cluster centroids (like in K-Means) and the cost of violating the pairwise constraints (must-link and cannot-link constraints between data patterns).

The MPC-KMeans algorithm (MPCKM) [15] is an extension of the PC-KMeans algorithm by proposing the incorporation of a metric learning directly into the clustering algorithm in a way that allows pairwise constraints to influence the metric learning process along with pairwise constraints. Basically the MPC-KMeans algorithm combines the objective function of the PC-KMeans algorithm with the learning of the distance metric.

II. Hierarchical algorithms create a hierarchical decomposition of a given set of data objects. A hierarchical algorithm can be agglomerative or divisive, based on how the hierarchical decomposition is formed. The agglomerative approach starts with each object forming a separate cluster and in each successive iteration merges the objects or clusters that are close to one another, until all of the clusters are merged into one, or until a termination condition holds. The divisive approach starts with all of the objects in the same cluster and in each successive iteration a cluster is split up into smaller clusters, until eventually each object is in one cluster, or until a termination condition holds. The single-link, average-link and complete-link [16] are examples of agglomerative algorithms.

III. Density-based algorithms [17] consider high-density regions in space as clusters, and objects in low-density regions as outliers or noise. Their general idea is to continue growing clusters as long as the density (number of objects or data points) in the "neighbourhood" exceeds some threshold.

IV. Grid-based algorithms [18] quantize the object space into a finite number of cells obtained by splitting each data feature into intervals. These cells form a grid structure. Clusters are formed by finding contiguous cells containing a minimum number of objects.

V. Spectral algorithms use the highest K eigenvalues to build a new representation of data. Then, a fast clustering algorithm, such as K-Means, is applied to perform clustering on the new representation.

The Normalized Cut algorithm [19] transforms the clustering algorithm into a weighted graph partitioning problem $G = (V, E)$, such that the vertices of the graph $V = \{v_1, \ldots, v_n\}$ correspond to the data patterns and the weights w_{ij} for each edge $E = \{e_{ij} : 1 < i < n-1, 2 < j < n, i < j\}$ correspond to the similarity between a pair of data patterns, and partitions the graph into K clusters.

VI. Model-based algorithms [20] assume a model for each cluster and find the best fit of the data to the given model. They locate clusters by constructing a density function that reflects the spatial distribution of the data points.