Engineering Applications of Artificial Intelligence 38 (2015) 1–13

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Evaluation of the performance of clustering algorithms for a high voltage industrial consumer

Ioannis Panapakidis, Minas Alexiadis, Grigoris Papagiannis *

Power Systems Laboratory, School of Electrical and Computer Engineering, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece

ARTICLE INFO

Article history:
Received 22 February 2014
Received in revised form
14 October 2014
Accepted 14 October 2014
Available online 7 November 2014

Keywords:
Demand side management
Load modeling
Load profiles
Conditional entropy minimization clustering
Unsupervised machine learning

ABSTRACT

Load profiling refers to a procedure which leads to the formulation of daily load curve clusters based on the similarity of the curves shapes. This paper focuses on the investigation of the consumption patterns of an existing high voltage industrial consumer. The profiling process involves stages like the normalization of the recorded load data, the utilization of pattern recognition algorithms, the selection of the appropriate validation scheme and the exploitation of the profiling findings. Certain improvements are proposed for each of these stages. More specifically, the most common algorithms of the related literature are implemented and a detailed investigation of their performance is presented. A new algorithm is proposed, presenting, in the majority of the cases, the best performance. Additionally, all the clustering validity indicators of the literature are considered to evaluate the clustering results. After the formulation of the load curve clusters, the load profiles are extracted and based on specific indices conclusions are drawn regarding the implementation of suitable demand side management schemes.

1. Introduction

Electricity consumer classification is a prominent scheme in the deregulated markets, since it can serve as the basis for designing suitable tariffs that directly reflect the generation costs. The consumers of the same class present similar daily demand patterns and different pricing policies can be applied, leading to increased profitability of retailers (Mahmoudi-Kohan et al., 2010). Load profiling is a tool that provides the required information about the demand behavior among various kinds of consumers, so that the aforementioned actions can be effectively implemented.

Given a set of metered load data, the task of grouping them in a sophisticated manner is a subject of many requirements (Panapakidis et al., 2013). First of all, a representative sample of load curves must be obtained when dealing with the load profiling of either an individual consumer or a group of consumers. After gathering the metered data, we need to specify the load condition. This refers to the periodic attributes of the demand. For example, the load profiling may consider only working days excluding holidays and weekends. In our work, we formulated nine yearly daily load data sets. Next, a suitable preprocessing of the data is necessary. The data are normalized in the appropriate range, since in the clustering operation we are concerned mainly with the load curve shape similarities. The load profiling process involves the proper selection of one or more clustering algorithms. Each algorithm has unique characteristics, such as input parameter requirements, speed and computational complexity. The purpose of the clustering procedure is the search of structures within a data sample. The formation of groups takes place, where the population within the same group shows more similar characteristics compared with the members of the other groups. The clustering procedure is data driven. In the majority of the applications there is no a priori knowledge about the relations between the data, their geometry or special attributes.

2. Literature survey and contributions

In most cases, the appropriate number of daily load curve clusters is unknown. Therefore a load profiling problem is formulated as an unsupervised learning task (Panapakidis et al., 2013). Clustering based load profiling is a two-stage process. In the first stage, each consumer is studied separately. His daily load curves are clustered and certain classes are formed. Each class is represented by the average daily load curve, which is actually the normalized load profile. In the second stage, a specific load profile is chosen for each consumer and the second clustering occurs that generates consumer classes (Tsekouras et al., 2007). Otherwise, the first stage clustering can be skipped and the load profile that represents the consumer can be an averaged load curve, which corresponds separately on weekdays and weekends (Chicco et al., 2004).

A number or researchers have proposed various clustering algorithms and the performance of the algorithms is tested by a set of adequacy measures, which are indications of the within-cluster

* Corresponding author. Tel.: +30 2310 996388; fax: +30 2310 996356.
E-mail addresses: ipanap@ee.auth.gr (I. Panapakidis), minalex@auth.gr (M. Alexiadis), grigoris@eng.auth.gr (G. Papagiannis).

http://dx.doi.org/10.1016/j.engappai.2014.10.013
0952-1976/

distances of the features that belong to the same cluster and the centroid (center) of it and the distances between the centroids of the different clusters. These researches can be generally categorized based on the type of the algorithm that they used and based on the characteristics of the data sets under study. The first category includes studies that involve:

- Partitional clustering algorithms, like the K-means, Weighted Fuzzy Average (WFA) K-means, Improved Weight Fuzzy Average (IWFA) K-means and modified "follow-the-leader" (Tsekouras et al., 2007; Chicco et al., 2004; Bidaki et al., 2010; Kohan et al., 2009; Cao et al., 2013; Ramos et al., 2013; Kwac et al., 2014).
- Hierarchical clustering algorithms, like the family of agglomerative algorithms and the COBWEB (Tsekouras et al., 2007; Chicco et al., 2004; Bidaki et al., 2010; Gerbec et al., 2002a, 2002b, 2003; Nizar et al., 2006; Hino et al., 2013; Notaristefano et al., 2013; Kwac et al., 2014).
- Fuzzy clustering algorithms, like the Fuzzy C-Means (FCM) (Tsekouras et al., 2007; Chicco et al., 2004; Bidaki et al., 2010; Chicco and Ilie, 2009; Anuar and Zakaria, 2012; Gerbec et al., 2002b; Zakaria and Lo, 2009; Iglesias and Kastner, 2013).
- Neural Network (NN) based clustering algorithms, like the Self-Organizing Map (SOM), Adaptive Vector Quantization (AVQ) trained NN and Hopfield NN (Tsekouras et al., 2007; Chicco et al., 2004; Bidaki et al., 2010; Rodrigues et al., 2003; Chicco and Akilimali, 2009; Chicco and Ilie, 2009; Räsänen et al., 2010; Figueiredo et al., 2005; López et al., 2011; Verdu et al., 2004; Wang et al., 2013).
- Algorithms that do not belong to the above categories, like the Expectation Maximization (EM), Support Vector Clustering (SVC), Renyi entropy (Nizar et al., 2006; Chicco and Akilimali, 2009; Chicco and Ilie, 2009), Ant Colony Clustering (Chicco et al., 2013) and a set of Subspace and Projection Clustering Methods (Piao et al., 2014).

The second category distinguishes among studies that involve load data sets from:

- Low voltage consumers (Räsänen et al., 2010; Figueiredo et al., 2005; López et al., 2011; Rodrigues et al., 2003; Nizar et al., 2006, Anuar and Zakaria, 2012; Hino et al., 2013; Cao et al., 2013; Iglesias and Kastner, 2013; Piao et al., 2014).
- Medium voltage consumers (Tsekouras et al., 2007; Chicco et al., 2004; Bidaki et al., 2010; Kohan et al., 2009; Gerbec et al., 2002a, 2002b, 2003; Chicco and Akilimali, 2009; Chicco and Ilie, 2009; Notaristefano et al., 2013; Chicco et al., 2013; Ramos et al., 2013; Wang et al., 2013).

In spite of the extended research in the field of load profiling, a more sophisticated segmentation of recorded load data is still required. The contribution of the present paper can be summarized as follows:

1. A detailed comparison of the most commonly used algorithms in the load profiling is performed. We also introduce a new type of clustering concept. The Minimum Entropy Clustering (minCEntropy) is an objective-function-oriented algorithm where the task is to minimize the conditional entropy within the clusters (Vinh and Epps, 2010). The algorithm has already shown robust performance in other clustering applications. In this study, we introduce a modified version of this algorithm, which includes hybridization with the K-means. The latter is used to provide the initial centroids of the minCEntropy. The superiority of the hybrid K-means/minCEntropy algorithm over the others already considered in the literature is demonstrated.
2. This is the first study that examines all the clustering validity indicators or adequacy measures of the literature used to evaluate the algorithms operation. An effectual partitioning of the metered samples should lead to low values of intra-connectivity and increased degree of inter-connectivity of the generated clusters. The clusters should be well-separated, so that a post-processing of the grouped data would be feasible and accurate. The performance of the different validity indicators for the different algorithms is investigated, but also their implementation in defining the optimal number of clusters.
3. So far in the literature, the clustering is only evaluated in a strictly mathematical manner (i.e. considering the values of the indicators). Several additional performance criteria are introduced in this paper, like the computation time of the algorithms, their complexity as expressed by the input parameters requirements, whether a parameter updating towards the simulations is needed, whether the clustering results are exploitable by a specific load profiling application and others.
4. While the majority of the related studies terminate their analysis in the comparison of the algorithms, we move a step forward to a load profiling application, providing a framework for the usage of the load profiling findings in the design of the appropriate Demand Side Management (DSM) measure for the consumer under study.

The majority of the respective researches deals with the profiling of the medium voltage (MV) consumers; the load profiling of high voltage (HV) industrial consumers have not been proposed so far in the literature (Chicco, 2012; Zhou et al., 2013). The analysis of the present study is focused on the daily load curves of an existing HV metal industry located in Greece. The HV consumers' consumption in Greece was 6.61 TWh in 2011, corresponding to the 12.63% of the total consumption of the Greek electric sector (Regulatory Authority for Energy, 2012). Due to the contractual DSM policies, HV industrial consumers are a group of special interest to the utilities. Since 2011 an ongoing discussion between the Public Power Corporation of Greece S.A. (PPC S.A.), which is the sole supplier of the HV industries, and the Regulatory Authority of Energy of Greece takes place, regarding tariffs offered to the HV customers. Load profiling of HV consumers can therefore help in the design and evaluation of sophisticated DSM measures and the analysis of the paper is oriented towards this goal.

3. Load profiling framework

3.1. Data collection

The data set used in this work is composed by nine subsets, corresponding to different years, between 2003 and 2011. To provide a general overview of the yearly demand, Table 1 registers the minimum, maximum and average load and daily consumed energy per year, together with the annual variation from the base year of 2003. The null values of Table 1 refer to temporary terminations of the industrial activity. It can be noticed that the maximum load continually increases until 2007. Then for the next two years the maximum annual load is reduced and finally in 2010 and 2011 rises again. The annual variation of the maximum load is below 10% while in 2011 the maximum load increase is 34.44%. This high increment of the demand is also recorded in the corresponding maximum daily consumed energy. Regarding the average yearly load, more fluctuations between the years are observed. For three years the average yearly load is reduced. This fact is reflected to the average daily consumed energy. Therefore the industry load does not follow a constant trend and Table 1 reveals that it is necessary to use an expanded data set to provide a more robust analysis for the patterns of electricity usage.