Advanced Engineering Informatics 28 (2014) 311–326

Contents lists available at ScienceDirect

Advanced Engineering Informatics

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

An unsupervised hierarchical clustering based heuristic algorithm for facilitated training of electricity consumption disaggregation systems

Farrokh Jazizadeh [a,*], Burcin Becerik-Gerber [b], Mario Berges [c], Lucio Soibelman [d]

[a] Sonny Astani Dept. of Civil and Environmental Engineering, Univ. of Southern California, KAP 217, 3620 South Vermont Ave., Los Angeles, CA 90089-2531, United States
[b] Sonny Astani Dept. of Civil and Environmental Engineering, Univ. of Southern California, KAP 224C, 3620 South Vermont Ave., Los Angeles, CA 90089-2531, United States
[c] Civil and Environmental Engineering Department, Carnegie Mellon University, Porter Hall 113, Pittsburgh, PA 15213-3890, United States
[d] Sonny Astani Dept. of Civil and Environmental Engineering, Univ. of Southern California, KAP 210A, 3620 South Vermont Ave., Los Angeles, CA 90089-2531, United States

ARTICLE INFO

Article history:
Received 14 February 2014
Received in revised form 6 August 2014
Accepted 11 September 2014
Available online 19 October 2014

Keywords:
Clustering
Unsupervised learning
Non-intrusive load monitoring
Training

ABSTRACT

Provision of training data sets is one of the core requirements for event-based supervised NILM (Non-Intrusive Load Monitoring) algorithms. Due to diversity in appliances' technologies, in-situ training by users is often required. This process might require continuous user-interaction to ensure that a high quality training data set is provided. Pre-populating a training data set could potentially reduce the need for user-system interaction. In this study, a heuristic unsupervised clustering algorithm is presented and evaluated to enable autonomous partitioning of appliances signature space (i.e. feature space) for applications in electricity consumption disaggregation. The algorithm is based on hierarchical clustering and uses the characteristics of a cluster binary tree to determine the distance threshold for pruning the tree without *a priori* information. The algorithm determines the partition of a feature space recursively to account for multi-scale nature of the binary cluster tree. Evaluation of the algorithm was carried out using metrics for accuracy and cluster quality (proposed in this study) on a fully labeled data set that was collected and processed in a real residential setting. The algorithm performance in accurate partitioning of the feature space and the effect of different feature extraction techniques were presented and discussed.

1. Introduction

Non-Intrusive Load Monitoring (NILM) as a low-cost alternative approach to individual load level monitoring has been the subject of numerous studies in the recent decades. NILM methodologies apply few sensing points (commonly at the circuit panel level) and use signal processing, pattern recognition, and inference algorithms for recognizing the underlying patterns of appliances' operational states and consequently the energy consumption associated with an operational schedule. In general, the NILM approaches could be categorized into two main classes: event-based and non-event based. In the existing event based NILM solutions, the load disaggregation problem is tackled by tracking the events on aggregated signal time series. Events are defined as variations in the time series caused by appliances' operational state changes such as turning on or off a light bulb. In recent years, non-event based NILM approaches have been also the subject of a few research studies, where mainly Hidden Markov Models (HMM) are used for decomposing aggregated power time series into constituent components (i.e., the power trace of individual contributors in the aggregated power draw) [1–4].

In event-based NILM approaches, detection of underlying state changes (associated with each event) is commonly defined as a classification (i.e., a supervised learning) problem, which requires provision of a training data set. Although, the application of generalized training data (using cross-building training process) has been the subject of a number of research studies, in most of the cases, upon installation of a NILM system, in-situ training has to be carried out due to the diversity in the design and manufacturing technologies used in different appliances, as well as various sources of noise in a data acquisition process. Considering these sources of complexity and the ad hoc nature of the problem, the training process needs to be continuous and requires supervision of a classification algorithm for a period of time to ensure that all the variations in the training data is introduced to the NILM system. This supervision calls for numerous user interactions with

* Corresponding author. Tel.: +1 213 400 2413.
E-mail addresses: jazizade@usc.edu (F. Jazizadeh), becerik@usc.edu (B. Becerik-Gerber), marioberges@cmu.edu (M. Berges), soibelman@usc.edu (L. Soibelman).

http://dx.doi.org/10.1016/j.aei.2014.09.004
1474-0346/

the NILM system, which is considered as one of the challenges in wide adoption of NILM systems.

To address these challenges and to facilitate the training process, one possible solution is to provide a pre-populated training data set, for which users can provide labels. Using the pre-populated training data set could potentially reduce the number of interactions, which in turn could improve user experience with the training process. Such training data sets provide the information about possible appliances operational states in a specific setting and thus provide the ground for smart communication with users (instead of communicating for all detected events). Pre-populating a training data set requires grouping the events' signatures (i.e., the signal characteristics in vicinity of events) into similar classes, which then could be labeled in the training process. Such a data set includes a number of examples for each class (e.g., the class of turn-on event of a television) and therefore, labeling one of those examples results in labeling the rest. This could be achieved by using clustering algorithms, which are used to group signatures into similar clusters, where the members of each cluster are similar to each other compared to the signatures in other clusters based on a predefined similarity metric. Although clustering is an unsupervised approach, for the majority of the clustering algorithms determining the representative number of clusters calls for *a priori* information. However, as noted, our objective is to reduce the challenges of the training process. The number of appliances' state changes, associated with the number of events, depends on the number of appliances and their operational states. Differences in number of appliances in each building, their internal components (e.g., a refrigerator could include compressor, defrosting module, and light fixture), and the fact that not all of the operational states (e.g., the refrigerator compressor operation and defrost operation) could be observed and detected by users compound the problem. Consequently, determination of the number of appliances state transitions (i.e., the number of clusters) is not a trivial task and it requires close monitoring of appliances by trained users. Furthermore, due to the ad-hoc nature of the signature space, determining any other generalized parameters (such as threshold values in stopping rules) for autonomous clustering could also be a challenging task. Accordingly, in this study, we propose a heuristic algorithm based on hierarchical clustering to achieve the objective of autonomous clustering of similar events' signatures, associated with appliances operational state transitions without the need for prior information.

The paper is structured as follows. First a research background covering NILM research, as well as clustering techniques, is presented in Section 2. Section 3 describes different components of our methodology in generating pre-populated training data sets, including our proposed heuristic algorithm. The experimental set up including the test bed and data acquisition for validation of the algorithm is presented in Section 4. The evaluation of algorithm performance, including the performance metric and the sensitivity analysis for exploring the effect of different features on the algorithm, is presented in Section 5. Section 6 concludes the paper by providing a summary of the study, followed by the future directions of the authors' research.

2. Research background

2.1. Event-based NILM research

Event-based NILM approaches comprise the main body of research since their introduction [5] three decades ago. A majority of the studies on NILM has focused on improvements to the methodologies, used for electricity disaggregation, through the introduction of different features [6] that represent appliances' operational characteristics to enhance the performance of pattern recognition algorithms. Features are the constituents of the events' signatures, which are the signal characteristics associated to events. Fig. 1 illustrates a short segment of real power time series and a number of events. Application of different features depends on signal resolution and consequently the data acquisition and processing system. Steady state features, that are commonly associated with low-resolution (e.g. 1 Hz) power metrics, are the features related to the segments of a power time series that are in the steady condition between the events (Fig. 1a).

Steady state real and reactive power metrics were used in Hart's seminal research [5] to improve the performance of the NILM system for appliances with a similar power draw. Since then, many studies used different combinations of steady state features. Real power, reactive power, power factor, RMS (Root Mean Square – used to identify the effective value of the alternating raw signal) current, and RMS voltage are the features that were used in some of the recent studies [7–9]. However, as illustrated in Fig. 1, the increase in signal resolution could potentially increase the information related to events and therefore various studies have focused on the application of the features associated with the high-resolution signals. These feature extraction methods have been coupled with different pattern recognition algorithms, such as nearest neighbor [10,11], neural networks [12,13], and Bayes classifiers [11]. Application of various feature extraction methods along with their pertinent methods for pattern recognition has been reviewed in [6].

As another approach, different sets of features were introduced in previous research, using generated electric noise as features for event detection and classification [14,10]. In [14], Patel et al. used a unique voltage transient noise, which is generated as a result of the operation of electromechanical components of a circuit (i.e., switching appliances on/off) and could be sensed at one outlet. The frequency component of the noise was used as a feature for pattern recognition. However, in [10], Gupta et al. pointed to the challenges such as the computational complexity and dependency of transient features on wiring and introduced the application of steady-state voltage noise features. Features are extracted from continuous electromagnetic interference (EMI) generated by appliances' switch mode power supply (SMPS). Modern appliances, with electronic components, generate high frequency EMI, which could be captured using high sampling frequencies about 500 kHz and higher. These features have been found to be transferable for a number of electronic appliances across different residential settings [10]. However, this approach could be used only for appliances that are equipped with SMPS components. Furthermore, the challenges related to the variation in appliances' manufacturing technologies and the assessment of temporal stability are yet to be explored.

Despite the above-mentioned improvements in the field of NILM, the methods for training data set provision remained undiscussed until recent years. Except for the sensor assisted approaches (for example [15–17]), Berges et al. [11,18] proposed a framework as a user-centered event-based NILM system to facilitate user interaction for training. In this framework, communication between a user and a NILM system has been facilitated through event detection and classification algorithms to continuously improve pattern recognition performance. However, continuous interaction between a user and a NILM system could potentially decrease the success of a NILM system set up. Accordingly, in this paper, we proposed a methodology for pre-populating training data sets through a heuristic clustering algorithm, explored the performance of the algorithm, and evaluated the effect of different feature extraction methods. The proposed algorithm enables unsupervised clustering to be utilized for facilitated training of an event based NILM approach.