

Semi-Automatic Labeling for Public Non-Intrusive Load Monitoring Datasets

Lucas Pereira and Nuno J. Nunes
Madeira Interactive Technologies Institute
Funchal, Portugal
lucas.pereira@m-iti.org, njn@uma.pt

Abstract—In this paper we present and evaluate a semi-automatic labeling prototype to enable the creation of fully labeled energy disaggregation datasets from sub-metered data. Our results advocate in favor of our approach and show that it is possible to extract individual appliance transitions with considerable precision, as long as the individual appliance information is present in the sub-metered data, and its resolution is high enough.

Keywords—NILM; datasets; event-detection; electric energy;

I. INTRODUCTION

Non-Intrusive Load Monitoring (NILM), also known as single point energy disaggregation, is a technology that aims at disaggregating and estimating the consumption of individual appliances by applying signal-processing and machine-learning (ML) techniques to the aggregated energy consumption measured from a single location in the household electric-grid.

Early research in this topic dates back to the 1980s, when George Hart from the Massachusetts Institute of Technology (MIT) introduced his prototype Non-Intrusive Appliance Load Monitor [1]. It was only recently that this technology gained renewed attention from researchers, in part due to the growing impact of electricity in the worldwide overall energy demand, and the consequent need to promote a more sustainable generation, distribution, and consumption of electric energy [2].

Despite all the potential and expectations of this technology, only recently there has been a serious effort to systematically evaluate the existing solutions. For example, the **formal classification of NILM research** according to two different approach categories, namely: i) *event-based* (EB), which include techniques that work by keeping track of every appliance state transition (e.g. TV turning *on* or *off*) using event detection and classification, assuming that the system was previously trained (e.g. [1], [3]) and ii) *non event-based* (NEB) where no previous knowledge of the existing appliances is assumed and the load disaggregation is done by means of techniques like Hidden Markov models or temporal motif mining techniques (e.g. [4], [5]).

However, the greatest step towards the generalization of NILM research findings was the **emergence of publicly available energy disaggregation datasets** targeting the evaluation and benchmark of the existing approaches.

In high-level terms, NILM datasets are collections of electric energy metrics taken from houses in real world scenarios containing measurements from the consumption of

the entire house (taken at the mains) and of the individual loads, i.e. ground-truth data, obtained either by measuring each load at the plug-level or the individual circuit where the load is connected.

Like NILM approaches, datasets are also categorized as event-based or non event-based, depending on which approaches they support. The major difference between the two lies mostly on the type of ground-truth data provided. More precisely, non event-based datasets provide the consumption of the individual appliances in the form of a time-series (e.g. 1 Hz) whereas event-based datasets provide localization information of each appliance transitions in the whole-house data (e.g. timestamp and appliance name).

To the best of our knowledge, to date, there are six public datasets created specifically for NILM evaluation (see TABLE I for an overview of these datasets), from which five target NEB approaches and only one is targeted at EB techniques.

TABLE I. OVERVIEW OF NILM SPECIFIC PUBLIC DATASETS

Dataset	Available data and resolution	Approach	
		EB	NEB
REDD [4]	Active power at 1 Hz; Individual Circuit and Individual appliance every 3 - 4 seconds	✗	✓
AMPds [6]	Active, individual appliance and individual circuit every 1 minute	✗	✓
UK-Dale [7]	Current and voltage at 16 kHz; active, reactive and voltage RMS at 1 Hz; active and individual appliance every 6 seconds	✗	✓
iAWE [8]	Current, voltage, active, reactive, apparent and individual appliance and individual circuit at 1 Hz	✗	✓
ECO [9]	Current, voltage, active power and individual appliance at 1 Hz	✗	✓
BLUED [10]	Current and voltage at 12 kHz; active and reactive at 60 Hz; power events list (timestamp and label)	✓	✗

In this paper we argue that the main reason behind the lack of EB datasets is that the actual labeling process still relies on a heavy, lengthy, and error-prone manual inspection of the whole dataset. Thus preventing the emergence of fully labeled datasets.

In fact, according to TABLE I, there is a visible possibility of creating EB datasets from NEB datasets since they all provide information about the individual loads in their ground-truth data (e.g. most individual circuits contain timestamps and power readings for a single appliance), that only have to be extracted and mapped back to the whole house data.

Against this background, and considering the importance of having fully labeled datasets to enable fair comparisons and benchmarks between the different approaches, we propose a

method for creating fully labeled energy disaggregation datasets, following in a semi-automatic approach that can be generalized to other datasets in different domains.

We refer to it as *semi-automatic labeling*, in a sense that signal-processing algorithms are used to detect individual appliance transitions information from the ground-truth data, which are then validated by the end-user to guarantee its correctness.

II. SEMI-AUTOMATIC LABELING PROTOTYPE

The developed prototype consists of two main modules that we refer to as backend and frontend. A general overview of this approach is shown in Fig. 1 below.

Starting from the left hand side, the ground-truth data is loaded from the existing datasets and that data is fed to an event detector algorithm (*arrow 1*). Next, each detected transition is stored in a local database with a respective timestamp and label (*arrow 2*).

Lastly, a graphical user interface is used to load the automatically detected power events and provide the end-user with the opportunity of supervising the labeling process.

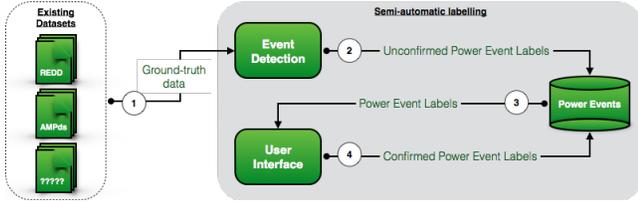


Fig. 1. General overview of the semi-automatic labeling prototype

In the backend, event detection algorithms are executed against each individual ground-truth circuit using different parameter combinations to understand which ones can provide better detection results. In the current prototype we are using the event detector described in our own previous work [11].

This algorithm works with a sliding window (*detection window*) that is used to calculate the likelihood of a change in mean to occur at a given sample and a second sliding window, called *extraction window* that is used to find the peak values of the likelihood test. The detection window $[i, k]$ is composed by two separate windows, $[i, j]$ and $[j, k]$, *pre-event* and *post-event* respectively.

For each sample in the power signal the likelihood of a power change occurring at that instant is given by (1), where $\mu[i, j]$ and $\mu[j, k]$ are the mean of the *pre-event* and *post-event* windows respectively, $\sigma[i, k]$ is the standard deviation of the *detection window* and $P(x)$ is the power value of the x^{th} sample.

$$l(x) = \frac{\mu_{[i,j]} - \mu_{[j,k]}}{\sigma_{[i,k]}^2} \times \left| P(x) - \frac{\mu_{[i,j]} + \mu_{[j,k]}}{2} \right| \quad (1)$$

In the frontend, end-users are presented with a visual representation of the events as shown in Fig. 2. This offers them the possibility of supervising the labeling process.

In the current prototype the users can confirm, edit and delete existing labels; end-users can also create new labels by clicking in the desired power measurement.

III. PROTOTYPE EVALUATION

Our prototype was evaluated against one week of data from the REDD and the AMPDs public datasets. Here we present the evaluation procedure and the obtained results.

A. Procedure

Ultimately, our goal is to find the best event detection parameter combination for each appliance, such that the number of correctly detected events in each circuit is maximized and consequently the need for user intervention minimized.

To this end, we tested several parameter combinations against the weekly data of each individual appliance and selected those that best served each of them. For selecting the best parameter combination we used the formulae in (2).

$$best_params = arg\ min_{params} [\sum_{d=1}^D |ed - ee| * w] \quad (2)$$

Where ed is the number of detected events, ee is number of expected events (i.e. ground-truth), d is one day, D is the total number of days and w is a weighing parameter in case we wish to have different weights for the missed events i.e. False Negatives (FN) and the wrong detections i.e. False Positives (FP). w is given by equation (3).

$$w = \begin{cases} x, & \text{if } (ed - ee) < 0 \\ 1 - x, & \text{if } (ed - ee) > 0 \end{cases} \quad (3)$$

Next we present the results of each test, according to the following metrics: True Positives (**TP**); False Positives (**FP**); False Negatives (**FN**); Precision (**Pre.**), i.e. the proportion of TP against all the obtained positive results (TP + FP); Sensitivity (**Sen.**), i.e. the proportion of TP against the expected positive results (TP + FN). Here w is set to 0.5, i.e., FN and FP will have the same impact in the final results.

B. REDD dataset

The REDD dataset, released in 2011, contains both whole house (1 Hz) and sub-metered active power measurements (one sample every 3 seconds) from six households in the United States.

For this test we selected one week of data (from April 18th 2011 to April 24th 2011) from household number two, consisting of whole house consumption for phases A and B as well as sub-metered consumption information for 8 individual circuits.

1) Labeling results

In the particular case of the REDD dataset (TABLE II), we immediately notice the high precision values, indicating that for each power event that is detected there is a high probability that it is an actual power event (0.97 probability). However, the low sensitivity (0.75) indicates that there are still a considerable number of events (more precisely 25%) to be detected.

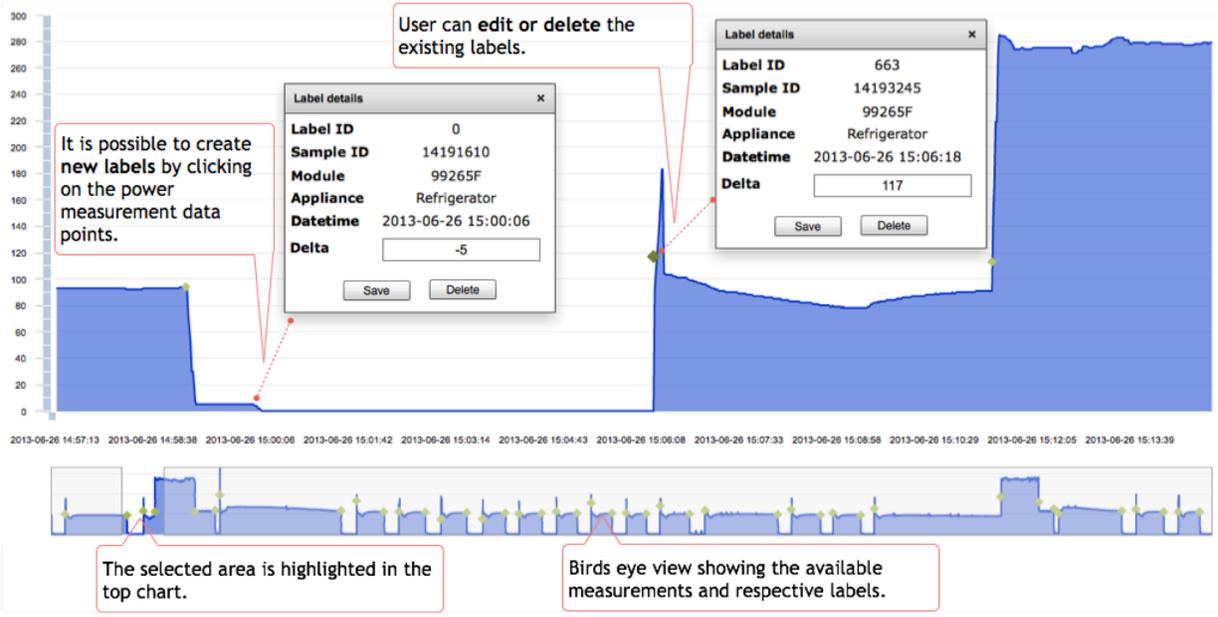


Fig. 2. Semi-automatic labeling prototype user interface

TABLE II. REDD DATASET: ONE-WEEK DATA RESULTS

Appliance	GT Events	Event Detection Results				
		TP	FP	FN	Pre	Sen
Dishwasher	59	48	0	11	1	0.81
Disposal	8	5	0	3	1	0.63
Refrigerator	610	515	0	95	1	0.84
Kitchen outlets 1	26	17	1	9	0.94	0.65
Kitchen outlets 2	426	226	0	200	1	0.53
Lighting	63	62	3	1	0.95	0.98
Microwave	48	48	21	0	0.7	1
Oven	18	18	0	0	1	1
Total	1258	939	25	319	0.97	0.75

The low sensitivity values are particularly relevant in the case of the refrigerator (~30% of the total misses) and the kitchen outlets circuit number 2 (~63% of the missed events).

A more in-depth look at the individual results clarifies the reason for the high number of FN in the refrigerator (about 15% of the events are missed) and the kitchen outlets # 2 (47% of the events are missed), which are directly related to the fact that most of the events in these circuits happen in short sequences of less than 3 seconds making it therefore hard to represent at this resolution, and consequently to be properly detected.

In short, in these two situations what happens is that the event detection algorithm is only able to detect one of the transitions in short sequence (either the *on* of the *off*). This effect is shown in Fig. 3, where it is possible to see the detection results for 35 minutes of the kitchen outlets # 2.

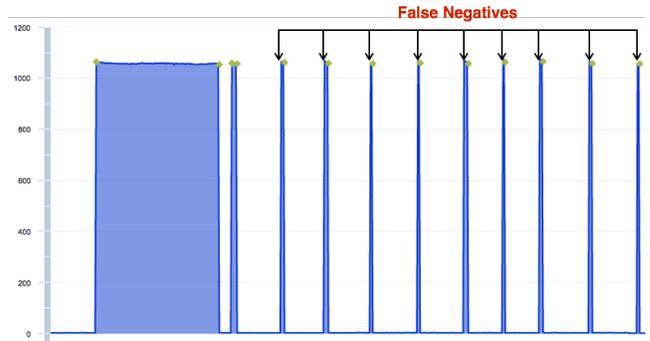


Fig. 3. 35 minutes of sub-metered data from the Kitchen outlets circuit # 2.

In order to better understand this effect we have isolated, in the whole house consumption data at 1 Hz, the periods for which we had the most false negatives and tested different combinations of the event detector parameters (see Fig. 4).

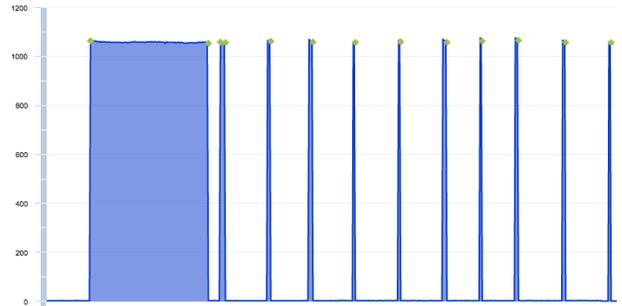


Fig. 4. 25 minutes of whole house consumption at 1 Hz

In this particular situation the algorithm managed to detect all the transitions, which is in accordance to our suspects that

using only three measurements per second is clearly not enough to detect events in close succession.

C. AMPds dataset

Released in 2013, the AMPds is public dataset containing two years of aggregate and sub-metered data (at 1 minute intervals) for one household in Canada.

For this test we selected one week of data (from April 1st 2012 to April 7th 2012) including consumption from phases A and B and five sub-metered appliances.

1) Labeling results

In the case of the AMPds (TABLE III.) the results of the event detection seemed very promising at a first glance (0.95 precision and sensitivity).

Yet, these high values are influenced by the perfect performance of the algorithm when applied to the refrigerator. In fact, results are quite poor for some particular appliances, namely the dishwasher and the clothes dryer (0.62 and 0.67 sensitivity), and the clothes washer that yield a considerably high number of FP, and consequently a very low precision (0.29) despite the perfect value for sensitivity.

TABLE III. AMDDS DATASET: ONE-WEEK DATA RESULTS

Appliance	GT Events	Event Detection Results				
		TP	FP	FN	Pre	Sen
Dishwasher	32	20	0	12	1	0.63
Clothes dryer	55	37	3	18	0.93	0.67
Clothes washer	12	12	30	0	0.29	1
Fridge	547	547	1	0	1	1
Wall-oven	5	4	2	1	0.67	0.8
Total	651	620	36	31	0.95	0.95

Looking at the data for each individual appliance it is possible to understand the main reasons behind these results. For example, Fig. 5 shows a full cycle of the clothes dryer and it is easily noticed that the event detector was not able to detect any two consecutive events, due to the low resolution of the dataset – 1 point per minute – (thus the high number of FN and subsequent low sensitivity).

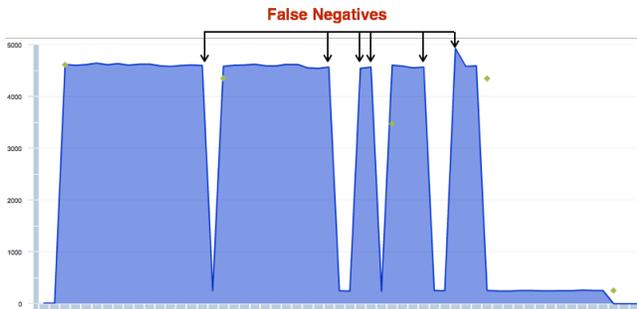


Fig. 5. A full clothes dryer cycle (Around 55 minutes)

These results highlight the importance of data resolution in event-based approaches. In particular, when considering

cycling appliances (like a clothes washer) where the several transitions happening in very short period of time are lost during the successive averaging steps.

IV. CONCLUSION AND FUTURE DIRECTIONS

In this paper we have presented and evaluated an approach that we believe can help produce better datasets by attempting to provide labels to the data in a semi-automatic fashion, thus alleviating researchers from the burden of doing this process manually.

Our initial results clearly advocate in favor of such a solution, especially if we consider the enormous amounts of data that are generated by energy monitors. Furthermore, our results indicate that despite the considerable number of FP and FN in some particular appliances, when the sub-metered data is available it is possible to automatically find and label most of the transitions in the dataset with minimum user intervention.

On the other hand, we have learned that there are still some interesting challenges that must be addressed in future versions. In particular, the automatic event detection process that has proven to degrade rapidly when the individual circuit data does not have a resolution that is high enough to detect power events that happen in quick succession.

In future iterations of this work we will add and evaluate different detection algorithms so that it is possible to handle sub-metered data with different time resolutions. Furthermore, we will also evaluate our approach in other datasets for longer periods of time so that the effectiveness and feasibility of this approach can be verified.

ACKNOWLEDGMENT

This research was sponsored by the FCT doctoral grant SFRH/DB/77856/2011 and the LARSyS project PEst-OE/EEI/LA0009/2015.

REFERENCES

- [1] G. W. Hart, "Prototype Nonintrusive Appliance Load Monitor," MIT Energy Laboratory Technical Report, and Electric Power Research Institute Technical Report, Sep. 1985.
- [2] Pacala and R. Socolow, "Stabilization wedges: solving the climate problem for the next 50 years with current technologies," *Science*, vol. 305, no. 5686, pp. 968–972, 2004.
- [3] M. Berges, E. Goldman, H. Matthews, L. Soibelman, and K. Anderson, "User-Centered Nonintrusive Electricity Load Monitoring for Residential Buildings," *J. Comput. Civ. Eng.*, vol. 25, no. 6, pp. 471–480, 2011.
- [4] Z. Kolter and J. Matthew, "REDD: A public data set for energy disaggregation research," SustKDD '11, San Diego, CA, USA.
- [5] H. Shao, M. Marwah, and N. Ramakrishnan, "A Temporal Motif Mining Approach to Unsupervised Energy Disaggregation: Applications to Residential and Commercial Buildings," AAAI '13, Bellevue, WA, USA.
- [6] S. Makonin, F. Popowich, L. Bartram, B. Gill, and I. V. Bajic, "AMPds: A Public Dataset for Load Disaggregation and Eco-Feedback Research," EPEC '13, Halifax, NS, Canada.
- [7] J. Kelly and W. Knottenbelt, "'UK-DALE': A dataset recording UK Domestic Appliance-Level Electricity demand and whole-house demand," *arXiv:1404.0284 [cs]*, 2014.
- [8] N. Batra, M. Gulati, A. Singh, and M. B. Srivastava, "It's Different: Insights into Home Energy Consumption in India," in BuildSys '13, Rome, Italy.
- [9] C. Beckel, K. Kleiminger, R. Cicchetti, T. Staake, and S. Santini, "The ECO Data Set and the Performance of Non-Intrusive Load Monitoring Algorithms," BuildSys '14, Memphis, TN, USA.
- [10] K. Anderson, A. Ocleanu, D. Benitez, D. Carlson, A. Rowe, and M. Berges, "BLUED: A Fully Labeled Public Dataset for Event-Based Non-Intrusive Load Monitoring Research," SustKDD '12, Beijing, China.
- [11] L. Pereira, F. Quintal, R. Gonçalves, and N. J. Nunes, "SustData: A Public Dataset for ICT4S Electric Energy Research," ICT4S '14, Stockholm, Sweden.