NILMTK: An Open Source Toolkit for Non-intrusive Load Monitoring

[Extended Abstract]

Nipun Batra[1], Jack Kelly[2], Oliver Parson[3], Haimonti Dutta[4], William Knottenbelt[2], Alex Rogers[3], Amarjeet Singh[1], Mani Srivastava[5]

[1]Indraprastha Institute of Information Technology Delhi, India {nipunb, amarjeet}@iiitd.ac.in
[2] Imperial College London {jack.kelly, w.knottenbelt}@imperial.ac.uk
[3] University of Southampton {osp, acr}@ecs.soton.ac.uk
[4] CCLS Columbia {haimonti@ccls.columbia.edu}
[5] UCLA {mbs@ucla.edu}

***Abstract*—Non-intrusive load monitoring, or energy disaggregation, aims to separate household energy consumption data collected from a single point of measurement into appliance-level consumption data. In recent years, the field has rapidly expanded due to increased interest as national deployments of smart meters have begun in many countries. However, empirically comparing disaggregation algorithms is currently virtually impossible. This is due to the different data sets used, the lack of reference implementations of these algorithms and the variety of accuracy metrics employed. To address this challenge, we present the Non-intrusive Load Monitoring Toolkit (NILMTK); an open source toolkit designed specifically to enable the comparison of energy disaggregation algorithms in a reproducible manner. This work is the first research to compare multiple disaggregation approaches across multiple publicly available data sets. Our toolkit includes parsers for a range of existing data sets, a set of statistics for describing data sets, two reference benchmark disaggregation algorithms and a suite of accuracy metrics. We demonstrate the range of reproducible analyses which are made possible by our toolkit, including the analysis of six publicly available data sets and the evaluation of both benchmark disaggregation algorithms across such data sets. This is a summary of a full paper in-press at ACM e-Energy 2014 [1].**

I. Introduction

Non-intrusive load monitoring (NILM), or energy disaggregation, aims to break down a household's aggregate electricity consumption into individual appliances [3]. As such, informing a household's occupants of how much energy each appliance consumes empowers them to take steps towards reducing their energy consumption [2]. However, three core obstacles currently prevent the direct comparison of state-of-the-art approaches, and as a result may be impeding progress within the field. To the best of our knowledge, each contribution to date has only been evaluated on a single data set and consequently it is hard to assess whether such approaches generalise to new households. Furthermore, many researchers sub-sample data sets to select specific households, appliances and time periods, making experimental results more difficult to reproduce. Second, newly proposed approaches are rarely compared against the same benchmark algorithms, further increasing the difficulty in empirical comparisons of performance between different publications. Moreover, the lack of reference implementations of these state-of-the-art algorithms often leads to the reimplementation of such approaches. Third, many papers target different use cases for NILM and therefore the accuracy of their proposed approaches are evaluated using a different set of performance metrics. As a result the numerical performance calculated by such metrics cannot be compared between any two papers. These three obstacles have led to the proposal of successive extensions to state-of-the-art algorithms, while a comparison between such approaches remains impossible.

Against this background, we propose NILMTK[1]; an open source toolkit designed specifically to enable easy access to and comparative analysis of energy disaggregation algorithms across diverse data sets. NILMTK provides a complete pipeline from data sets to accuracy metrics, thereby lowering the entry barrier for researchers to implement a new algorithm and compare its performance against the current state of the art. The contributions of NILMTK are summarised as follows:

- We propose NILMTK-DF (data format), the standard energy disaggregation data structure used by our toolkit. NILMTK-DF is modelled loosely on the REDD data set format [5] to allow easy adoption within the community. Furthermore, we provide parsers from six existing data sets into our proposed NILMTK-DF format.
- We provide statistical and diagnostic functions which provide a detailed understanding of each data set. We also provide preprocessing functions for mitigating common challenges with NILM data sets.
- We provide implementations of two benchmark disaggregation algorithms: first an approach based on combinatorial optimisation [3], and second an approach based on the factorial hidden Markov model [5], [4]. We demonstrate the ease by which NILMTK allows the comparison of these algorithms across a range of existing data sets, and present results of their performance.
- We present a suite of accuracy metrics which enables the evaluation of any disaggregation algorithm compatible with NILMTK. This allows disaggregation algorithms to be evaluated for a range of use cases.

[1]Code: http://github.com/nilmtk/nilmtk (v0.1.0 was used for this paper.)

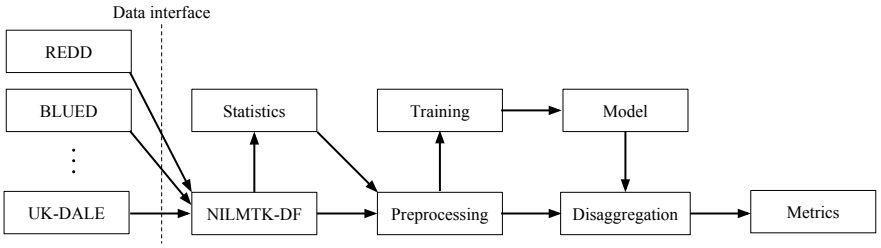

Fig. 1: NILMTK pipeline. At each stage of the pipeline, results and data can be stored to or loaded from disk.

II. NILMTK

We designed NILMTK with two core use cases in mind. First, it should enable the analysis of existing data sets and algorithms. Second, it should provide a simple interface for the addition of new data sets and algorithms. To do so, we implemented NILMTK in Python due to the availability of a vast set of libraries supporting both machine learning research and the deployment of such research as web applications

Figure 1 presents the NILMTK pipeline from the import of data sets to the evaluation of various disaggregation algorithms over various metrics. We now discuss each module of the pipeline in the remainder of this section.

A. Data Format

We propose NILMTK-DF; a common data set format inspired by the REDD format [5], into which existing data sets can be converted. NILMTK currently includes importers for the following six data sets: REDD, Smart*, Pecan Street, iAWE, AMPds and UK-DALE. After import, the data resides in our NILMTK-DF in-memory data structure, which is used throughout the NILMTK pipeline. At multiple stages in the pipeline, data can be saved or loaded from disk in either CSV or HDF5 format. In addition, NILMTK allows rich metadata to be associated with a household, appliance or meter. For example, NILMTK can store the parameters measured by each meter (e.g. reactive power, real power), the mains wiring defining the meter hierarchy (useful if a single appliance is measured at the appliance, circuit and aggregate levels), etc.

B. Data Set Functions

NILMTK provides the following functions for data set diagnostics, statistics and preprocessing:

Detect gaps: Identifies pairs of consecutive samples where the time between them is larger than a predefined threshold.

Dropout rate: The total number of recorded samples, divided by the number of expected samples.

Dropout rate (ignoring gaps): Removes large gaps where the sensor is off and calculates the dropout rate.

Up-time: The total time which a sensor was recording data.

Diagnose: Checks for all the issues listed above.

Proportion of energy sub-metered: Quantifies the proportion of total energy measured by sub-metered channels.

Downsample: Down-samples data sets to a specified frequency using aggregation functions such as mean and median.

Voltage normalisation: Normalises power demands to take into account fluctuations in mains voltage.

Top-k appliances: Identifies the top-k energy consuming appliances.

NILMTK also provides preprocessing functions for fixing other common issues with these data sets, such as: (i) interpolating small periods of missing data when appliance sensors did not report readings, (ii) filtering out implausible values (such as readings where observed voltage is more than twice the rated voltage) and (iii) filtering out appliance data when mains data is missing. A detailed account of preprocessing and statistics functions supported by NILMTK can be found in the online documentation.[2]

C. Training and Disaggregation Algorithms

NILMTK provides implementations of two common benchmark disaggregation algorithms:

Combinatorial Optimisation: Finds the combination of appliance states which minimises the difference between the sum of the predicted appliance power and the observed aggregate power, subject to a set of appliance models [3]. However, since each time slice is considered as a separate optimisation problem, each time slice is assumed to be independent.

Factorial Hidden Markov Model: Models each appliance using a hidden Markov model, and therefore this approach models a household as a factorial hidden Markov model [5], [4]. Since the operational states of each appliance are modelled using a Markov chain, the dependence between time slices is modelled by this approach.

For algorithms such as FHMMs, it is necessary to model the relationships amongst consecutive samples. Thus, NILMTK provides facilities for dividing data into continuous sets for training and testing.

D. Appliance Model Import and Export

Many approaches require sub-metered power data to be collected for training purposes from the same household in which disaggregation is to be performed. However, such data is costly and intrusive to collect, and therefore is unlikely to be available in a large-scale deployment of a NILM system. As a result, recent research has proposed training methods which do not require sub-metered power data to be collected from each household [4], [6]. To provide a clear interface

[2]Documentation: http://nilmtk.github.io/nilmtk/

Data set	Number of appliances	Percentage energy sub-metered	Dropout rate (percent) ignoring gaps	Mains up-time per house (days)	Percentage up-time
REDD	9, 16, 23	58, 71, 89	0, 10, 16	4, 18, 19	8, 40, 79
Smart*	25	86	0	88	96
Pecan Street	13, 14, 22	75, 87, 150	0, 0, 0	7, 7, 7	100, 100, 100
AMPds	20	97	0	364	100
iAWE	10	48	8	47	93
UK-DALE	4, 12, 53	19, 48, 82	0, 7, 22	36, 102, 470	73, 84, 100

TABLE I: Summary of data set results calculated by the diagnostic and statistical functions in NILMTK. Each cell contains the minimum, median and maximum values across the set of houses. AMPds, Smart* and iAWE each contain just a single house, hence these rows have a single number per cell.

between training and disaggregation algorithms, NILMTK provides a *model* module which encapsulates the results of the training module required by the disaggregation module. Each implementation of the module must provide import and export functions to interface with a JSON file for persistent model storage. NILMTK currently includes importers and exporters for the described FHMM and CO approaches.

E. Accuracy Metrics

NILMTK provides the following set of metrics which combines both general and energy disaggregation metrics:

1) Error in total energy assigned
2) Fraction of total energy assigned correctly
3) Normalised error in assigned power
4) RMS error in assigned power
5) Confusion matrix
6) True/False positives, True/False negatives
7) True/False positive rate
8) Precision, Recall
9) F-score
10) Hamming loss

III. Evaluation

We now demonstrate several examples of the rich analyses supported by NILMTK. First, we diagnose some common (and inevitable) issues in a selection of data sets. Second, we show various patterns of appliance usage. Finally, we present summary performance results of the two benchmark algorithms across six data sets using a number of accuracy metrics.

A. Data Set Diagnostics

Table I shows a selection of diagnostic and statistical functions (defined in Section II-B) computed by NILMTK across six public data sets. The table illustrates that AMPds used a robust recording platform because it has a percentage up-time of 100%, a dropout rate of zero and 97% of the energy recorded by the mains channel was captured by the sub-meters. Similarly, Pecan Street has an up-time of 100% and zero dropout rate. However, two homes in the Pecan Street data registered a proportion of energy sub-metered of over 100%. This indicates that some overlap exists between the metered

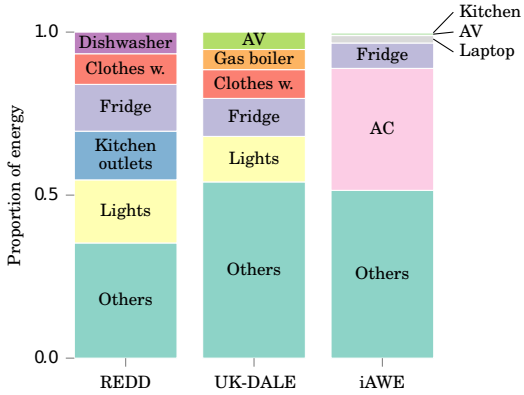

Fig. 2: Top 5 appliances in terms of the proportion of the total energy used in a single house (house 1) in each of REDD (USA), iAWE (India) and UK-DALE.

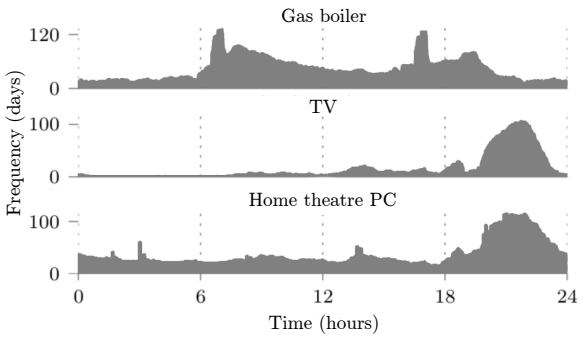

Fig. 3: Daily appliance usage histograms of three appliances over 120 days from UK-DALE house 1.

channels, and as a result some appliances are metered by multiple channels. This illustrates the importance of data set metadata (proposed as part of NILMTK-DF in Section II-A) describing the basic mains wiring.

B. Data Set Statistics

Figure 2 shows how the proportion of energy use per appliance varies between countries. It can seen that the REDD and UK-DALE households share some similarities in the breakdown of household energy consumption. In contrast, the iAWE house shows a vastly different energy breakdown, where the two air conditioning units account for almost half of the household's energy consumption. Figure 3 shows histograms of usage patterns for three appliances over an average day, showing similarities between groups of appliances. The usage patterns of the TV and Home theatre PC are similar because the Home theatre PC is the only video source for the TV. In contrast, the boiler's usage pattern is due to household's occupancy pattern and hot water timer in mornings and evenings.

C. Disaggregation Across Data Sets

We now compare the disaggregation results across the first house of six publicly available data sets. Since all the data sets

Data set	NEP		FTE		F-score	
	CO	FHMM	CO	FHMM	CO	FHMM
REDD	1.61	1.35	0.77	0.83	0.31	0.31
Smart*	3.10	2.71	0.50	0.66	0.53	0.61
Pecan Street	0.68	0.75	0.99	0.87	0.77	0.77
AMPds	2.23	0.96	0.44	0.84	0.55	0.71
iAWE	0.91	0.91	0.89	0.89	0.73	0.73
UK-DALE	3.66	3.67	0.81	0.80	0.38	0.38

TABLE II: CO and FHMM performance across six data sets

were collected over different durations, we used the first half of the samples for training and the remaining half for disaggregation across all data sets. Further, we preprocessed the REDD, UK-DALE, Smart* and iAWE data sets to 1 minute frequency using the down-sampling filter (Section II-B) to account for different aggregate and mains data sampling frequencies and compensating for intermittent lost data packets. The small gaps in REDD, UK-DALE, SMART* and iAWE were interpolated, while the time periods where either the mains data or appliance data were missing were ignored. AMPds and the Pecan Street data did not require any preprocessing.

Since both CO and FHMM have exponential computational complexity in the number of appliances, we model only those appliances whose total energy contribution was greater than 5%. Across all the data sets, the appliances which contribute more than 5% of the aggregate include HVAC appliances such as the air conditioner and electric heating, and appliances which are used throughout the day such as the fridge. We model all appliances using two states (on and off) across our analyses, although it should be noted that any number of states could be used. However, our experiments are intended to demonstrate a fair comparison of the benchmark algorithms, rather than a fully optimised version of either approach. We compare the disaggregation performance of CO and FHMM across the following three metrics defined in Section II-E: (i) fraction of total energy assigned correctly (FTE), (ii) normalised error in assigned power (NEP) and (iii) F-score. These metrics were chosen because they have been used most often in prior NILM work. Preferable performance is indicated by a low NEP and a high FTE and F-score.

Table II summarises the results of the two algorithms across the six data sets. It can be observed that FHMM performance is superior to CO performance across the three metrics for REDD, Smart* and AMPds. This confirms the theoretical foundations proposed by Hart [3]; that CO is highly sensitive to small variations in the aggregate load. The FHMM approach overcomes these shortcomings by considering an associated transition probability between the different states of an appliance. However, it can be seen that CO performance is similar to FHMM performance in iAWE, Pecan Street and UK-DALE across all metrics. This is likely due to the fact that very few appliances contribute more than 5% of the household aggregate load in the selected households in these data sets. For instance, space heating contributes very significantly (about 60% for a single air conditioner which has a power draw of 2.7 kW in the Pecan Street house and about 35% across two air conditioners having a power draw of 1.8 kW and 1.6 kW respectively in iAWE). As a result, these appliances are easier to disaggregate by both algorithms, owing to their relatively high power demand in comparison to appliances such as electronics and lighting. In the UK-DALE house the washing machine was one of the appliances contributing more than 5% of the household aggregate load, which brought down overall metrics across both approaches.

IV. Conclusions and future work

In this paper, we proposed NILMTK; the first open source toolkit designed to allow empirical comparisons to be made between energy disaggregation algorithms across multiple data sets. The toolkit defines a common data format, NILMTK-DF, and includes parsers from six publicly available data sets to NILMTK-DF. The toolkit further facilitates the calculation of data set statistics, diagnosing problems and mitigating them via preprocessing functions. In addition, the toolkit includes implementations of two benchmark disaggregation algorithms based on combinatorial optimisation and the factorial hidden Markov model. Finally, NILMTK includes implementations of a set of performance metrics which will enable future research to directly compare disaggregation approaches through a common set of accuracy measures. We demonstrated several analyses facilitated by NILMTK including: use of statistics functions to detect missing data, learning of appliance models from sub-metered data and comparing disaggregation algorithms across multiple data sets and accuracy metrics.

Future work will focus upon the addition of recently proposed training and disaggregation algorithms and data sets. For instance, larger data sets such as HES could also provide additional insight into disaggregation performance. In addition, recently proposed algorithms which do not require sub-metered power data for their unsupervised training could be compared against the current supervised algorithms. An additional direction for future work could be the use of a semantic wiki to maintain a communal schema for appliance metadata. Finally, the inclusion of a household simulator would allow disaggregation algorithms to be evaluated in a wider variety of settings than those represented by existing data sets.

References

[1] N. Batra, J. Kelly, O. Parson, H. Dutta, W. Knottenbelt, A. Rogers, A. Singh, and M. Srivastava. NILMTK: An Open Source Toolkit for Non-intrusive Load Monitoring. In *Fifth International Conference on Future Energy Systems (ACM e-Energy)*, Cambridge, UK, 2014.

[2] S. Darby. The effectiveness of feedback on energy consumption. *A Review for DEFRA of the Literature on Metering, Billing and direct Displays*, 2006.

[3] G. W. Hart. Nonintrusive appliance load monitoring. *Proceedings of the IEEE*, 80(12):1870–1891, 1992.

[4] H. Kim, M. Marwah, M. F. Arlitt, G. Lyon, and J. Han. Unsupervised Disaggregation of Low Frequency Power Measurements. In *Proceedings of 11th SIAM International Conference on Data Mining*, pages 747–758, Mesa, AZ, USA, 2011.

[5] J. Z. Kolter and M. J. Johnson. REDD: A public data set for energy disaggregation research. In *Proceedings of 1st KDD Workshop on Data Mining Applications in Sustainability*, San Diego, CA, USA, 2011.

[6] O. Parson, S. Ghosh, M. Weal, and A. Rogers. Non-intrusive load monitoring using prior models of general appliance types. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, pages 356–362, Toronto, ON, Canada, 2012.