Energy Efficiency (2015) 8:809–814
DOI 10.1007/s12053-014-9306-2

SHORT COMMUNICATION

Nonintrusive load monitoring (NILM) performance evaluation

A unified approach for accuracy reporting

Stephen Makonin · Fred Popowich

Received: 22 March 2014 / Accepted: 13 October 2014 / Published online: 31 October 2014

Abstract Nonintrusive load monitoring (NILM), sometimes referred to as load disaggregation, is the process of determining what loads or appliances are running in a house from analysis of the power signal of the whole-house power meter. As the popularity of NILM grows, we find that there is no consistent way the researchers are measuring and reporting accuracies. In this short communication, we present a unified approach that would allow for consistent accuracy testing.

Keywords Load disaggregation · Accuracy · Energy conservation · Smart grid

Introduction

Nonintrusive (appliance) load monitoring (NILM or NIALM) is the process of determining what loads or appliances are running in a house from analyzing the power signal of the whole-house power meter. NILM, which is sometimes called load disaggregation, can be used in systems to inform the occupants about how energy is used within a home without the need of purchasing additional power monitoring sensors. Once the occupants are informed about what appliances are running, and how much power these appliances consume, they can then make informed decisions about conserving power, whether motivated by economic or ecologic concerns (or both).

A review of NILM algorithms and research has led us and others (Kim et al. 2010; Zeifman and Roth 2011; Makonin 2014) to the conclusion that there is no consistent way to measure performance accuracy. Although some researchers still use the most basic forms of accuracy measure, there has been discussion concerning more sophisticated measurements. The most basic accuracy measure used by a majority of NILM researchers (e.g., Chang et al. 2010; Tsai and Lin 2012; Makonin et al. 2013) is defined as

$$\text{Acc.} = \frac{\text{correct matches}}{\text{total possible matches}} = \frac{\text{correct}}{\text{correct} + \text{incorrect}}. \quad (1)$$

Kim et al. (2010) point out that accuracy results are "very skewed because using an appliance is a relatively rare event .... appliances [that] are off will achieve high accuracy" (Table 1). Better accuracy performance measures must be considered. Expanding on our previous work (Makonin 2014), we present a

Electronic supplementary material The online version of this article (doi:10.1007/s12053-014-9306-2) contains supplementary material, which is available to authorized users.

S. Makonin (✉) · F. Popowich
Computing Science, Simon Fraser University, Burnaby, Canada
e-mail: smakonin@sfu.ca

F. Popowich
e-mail: popowich@sfu.ca

Table 1 Basic accuracy measures

Load	Acc (%)	TP	Inacc	APT	ITP	TN	FP	FN
Overall score	97.28	86,398	78.44	82,280	4111	474,850	9108	6610
Basement	96.34	5590	0.00	4710	879	44,942	947	973
Clothes dryer	99.35	647	0.10	647	0	51,461	43	300
Clothes washer	97.88	76	2.50	19	57	51,265	130	980
Dishwasher	98.80	863	4.52	845	17	50,959	335	294
Kitchen fridge	88.23	17,429	12.97	17,388	41	28,847	4587	1588
HVAC/furnace	99.90	52,376	35.67	50,893	1482	25	36	15
Garage	99.93	0	0.00	0	0	52,413	7	31
Heat pump	99.70	4622	22.27	4395	226	47,672	51	107
Home office	94.68	492	0.00	487	4	49,171	1173	1615
Ent/TV/DVD	95.43	4188	0.00	2782	1405	45,866	1762	636
Wall oven	99.79	115	0.42	114	0	52,229	37	71

unified approach that would allow for consistent accuracy testing amongst NILM and load disaggregation researchers.

The rest of our short communication is organized as follows. We first define data noise (Section Data noise), then discuss strategies using ground truth (Section Ground truth and bias). Next, we focus on classification accuracy testing (Section Classification accuracy) and estimation testing (Section Estimation accuracy). We end the discussion with a look at why researchers need to report accuracies with respect to both the overall performance and appliance-specific performance (Section Overall and appliance-specific accuracies). Finally, we demonstrate some of the issues that we discussed previously by examining the results from an experiment (Section Experiment example).

Data noise

Data noise can be understood as unexpected or unaccounted for anomalies that can appear in the stream of data that an algorithm analyzes. Noise can take a number of forms when looking at disaggregation. There can be readings that are missing, leaving gaps in a time series of data. There can be data streams that have timestamps that are out of sync. There can be corrupted data where data measurements within the reading are missing or measured wrongly due to sensor miscalculation or malfunction. Aside from miscalculation or malfunction, data can contain Gaussian noise due to small fluctuations in sensor/ADC (analog-to-digital converter) precision and the consumption of power by an appliance. Specifically for disaggregation, noise can be unmetered appliances that create large unexpected patterns of energy consumption. For our purpose, we define noise as the amount of power remaining in the observed aggregate power reading once the disaggregated appliance power readings (in ground truth) have been subtracted. Mathematically, defined as

$$\text{noise} = y_t - \sum_{m=1}^{M} y_t^{(m)}, \tag{2}$$

where y_t is the total ground truth or observed value at time t, M is the number of appliances, and $y_t^{(m)}$ is the ground truth power consumed at time t for appliance m.

Ground truth and bias

NILM researchers need to describe in detail the data they are using to build models, train, and test their NILM algorithms. If researchers are using data from publicly available datasets such as REDD (Kolter and Johnson 2011) or AMPds (Makonin et al. 2013), they need to discuss the method used to clean the data. For instance, discussing how they dealt with incomplete

or erroneous data and with different meters having different sample rates.

There also needs to be a clear statement on whether the testing included *noise* or was *denoised*. In denoised data, the whole-house power reading is equal to the summation of all appliance power readings—which we often refer to as the *unmetered* load or appliance. Using denoised data for testing will cause higher accuracies to be reported. Denoised data does not reflect a real-world application because there would be a significant amount of noise due to unmetered loads running in the home. Furthermore, what needs to be reported is the percentage of noise in each test. This *percent-noisy measure* (%-NM) would be calculated on the ground truth data as such:

$$\%-NM = \frac{\sum_{t=1}^{T} |y_t - \sum_{m=1}^{M} y_t^{(m)}|}{\sum_{t=1}^{T} y_t}, \tag{3}$$

where y_t is the aggregate observed current/power amount at time t and $y_t^{(m)}$ is the ground truth current/power amount for each appliance m to be disaggregated. For example, a denoised test would result in 0 %, whereas a %-NM of 0.40 would mean that 40 % of the aggregate observed current/power for the whole test was noise.

Finally, researchers should use standard methods to minimize any effects of bias. Bias occurs when some data used for training is also used for testing, and when present, results in the reporting of higher accuracies. A well-accepted method used by the data mining community to avoid bias is *10-fold cross-validation* (Liu and Motoda 1998, pp. 109). This simple method splits the ground truth data into ten subsets of size $\frac{n}{10}$. NILM algorithms can then be trained on nine of the subsets, and accuracy testing is performed on the excluded subset. This is repeated ten times (each time a different subset is used for testing) and the mean accuracy is then calculated and reported.

Classification accuracy

Researchers need to measure how accurately NILM algorithms can predict what appliance is running in each state. Classification accuracy measures, such as *f-score* (a.k.a. f-measure), are well suited for this task. F-score, often used in information retrieval and text/document classification, has also been used by NILM researchers (Figueiredo et al. 2012; Berges et al. 2010; Kim et al. 2010). It is the harmonic mean of *precision* and *recall*:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall},$$

$$precision = \frac{tp}{tp + fp}, \quad recall = \frac{tp}{tp + fn},$$

where *precision* is the positive predictive values and *recall* is the true positive rate or sensitivity, tp is true-positives (correctly predicted that the appliance was ON), fp is false-positives (predicted appliance was ON but was OFF), and fn is false-negatives (appliance was ON but was predicted OFF). Note these measures (tp, fp, fn) are accumulations over a given experimental time period. However, f-score is generally used for binary classification purposes.

Kim et al. (2010) showed how f-score could be modified to account for non-binary outcomes, such as a power signal (we call M-fscore). Their approach combined appliance state classification and power estimation accuracies together even though in many instances classification and estimation are two distinct functions of NILM algorithms. Combining classification and estimation hides important diagnostic information as to what parts of NILM algorithms have low accuracy. Furthermore, functions, such as classification and estimation, require a specific type of accuracy measure that is suited for measuring their performance. Matching function with accuracy measure provides more detailed diagnostic and performance information.

To calculate the accuracies of non-binary classifications, we now define *finite-state f-score* (FS-fscore). We introduce a partial penalization measure called *inaccurate portion of true-positives* ($inacc$) which converts the binary nature of tp into a discrete measure. The $inacc$ of a given experimental test is

$$inacc = \sum_{t=1}^{T} \frac{|\hat{x}_t^{(m)} - x_t^{(m)}|}{K^{(m)}}, \tag{4}$$

where $\hat{x}_t^{(m)}$ is the estimated state from appliance m at time t, $x_t^{(m)}$ is the ground truth state, and $K^{(m)}$ is the number of states for appliance m. In other words, we penalize based on the distance (or difference) of the estimated state and the ground truth state. Precision

and recall can now be redefined to account for these partial penalizations:

$$precision = \frac{tp - inacc}{tp + fp} \quad \text{and} \quad recall = \frac{tp - inacc}{tp + fn}. \quad (5)$$

The definition of f-score remains the same. A summation over all appliances M for each tp, $inacc$, fp, and fn (including a recalculation of precision, recall, and f-score) would allow for the overall classification accuracy of the experimental test to be reported.

Estimation accuracy

Accuracies based on power estimation also need to be reported to show how accurately the NILM algorithm can estimate how much power is being consumed compared to actual consumption. This is important because systems that use NILM need to report to the occupants what portion of the power bill can be attributed to each appliance. Additionally, when dealing with time-of-use billing (charging more per kWh at peak times), occupants need to know how much might have been saved if certain appliances (e.g., a clothes dryer) were not used during the peak period.

There are different accuracy measures that have been used to compare consumption estimation. Parson et al. (2012) has used root mean square error (RMSE) for reporting estimation accuracy. However, these measures are not normalized, and it is hard to compare how the disaggregation of one appliance performed over another. This becomes a bigger problem when you try to compare an appliance that consumes a large amount of power (e.g., heating) versus an appliance that consumes very little power (fridge).

Normalized disaggregation error (NDE) (Kolter and Jaakkola 2012; Parson et al. 2012; Dong et al. 2013) has also been used to measure the estimation accuracy of an appliance. With this measure, we would subtract the summation of all T estimations by the summation of all T ground truths. However, subtracting the summations would tend to report inflated accuracies because it is possible for errors to cancel each other out. For example, suppose we had an estimation of 2A and a ground truth of 0A at time t_1 and an estimation of 0A and a ground truth of 2A at time t_2, the NDE would be 0 % when in fact 100 % would be the correct error score. Kolter and Johnson (2011) and Johnson and Willsky (2013) estimation accuracy measure calculates the correct value of 0 % accurate (or 100 % error). We have chosen this estimation accuracy method to use and is defined as

$$Est.\ Acc. = 1 - \frac{\sum_{t=1}^{T}\sum_{m=1}^{M}|\hat{y}_t^{(m)} - y_t^{(m)}|}{2 \cdot \sum_{t=1}^{T}\sum_{m=1}^{M} y_t^{(m)}} \quad (6)$$

where T is the time sequence or number of disaggregated readings, M as the number of appliances, $\hat{y}_t^{(m)}$ is the estimated power consumed at time t for appliance m, and $y_t^{(m)}$ is the ground truth power consumed at time t for appliance m. This method allows for overall estimation accuracy reporting. By eliminating the summations over M, we can then report estimation accuracy for each appliance

$$Est.\ Acc.^{(m)} = 1 - \frac{\sum_{t=1}^{T}|\hat{y}_t^{(m)} - y_t^{(m)}|}{2 \cdot \sum_{t=1}^{T} y_t^{(m)}}. \quad (7)$$

Overall and appliance-specific accuracies

Both classification accuracy and estimation accuracy need to be reported in overall scores and appliance specific scores. Reporting how each appliance scores is important for identifying strengths and weaknesses of different NILM algorithms. With this more detailed accuracy information, one could imagine a system that would select different algorithms depending on the context (including specific history) of the disaggregation task. It is important also to keep in mind when reporting accuracies the result needs to be normalized. Normalized results allow the readers to understand the relative standings from one appliance to another and from each appliance to the overall accuracy. Finally, although more detailed information has its advantages, reporting specific scores for appliance states is not necessary because different makes/models of appliances will have a different number of states at different power levels.

Experiment example

We investigated how basic accuracy can be misleading by reporting high confidence numbers that do not accurately reflect inaccuracies in predicting rare

Table 2 Classification and estimation accuracy results

Load	F-Score (%)	M-fscore (%)	FS-fscore (%)	RMSE	NDE (%)	Est Acc (%)
Overall Score	91.66	87.30	91.58	4.9293	1.18	91.87
Basement	85.34	71.92	85.34	0.4134	6.86	83.06
Clothes Dryer	79.05	79.05	79.03	0.5750	5.79	96.17
Clothes Washer	12.05	3.01	11.65	0.4041	79.60	46.89
Dishwasher	73.29	71.82	72.91	0.5459	22.68	76.06
Kitchen Fridge	84.95	84.75	84.89	0.4480	12.75	82.57
HVAC/Furnace	99.95	97.12	99.88	0.2127	2.43	98.40
Garage	0.00	0.00	0.00	0.1102	92.14	46.17
Heat Pump	98.32	93.51	97.85	0.9178	0.50	97.05
Home Office	26.09	25.83	26.09	0.2501	24.32	34.53
Ent/TV/DVD	77.75	51.65	77.75	0.3018	9.02	68.55
Wall Oven	68.05	67.86	67.80	0.7503	6.03	75.67

events (Supplementary data). This would be the case for most loads that are sporadically used. We also show why modified f-score, which combines classification and estimation, is not a detailed enough measure. We used the more detailed AMPds (Makonin et al. 2013) rather than REDD (Kolter and Johnson 2011) to illustrate the issues with these different measurements, using our own NILM algorithm (Makonin et al. 2014). Current draw (I) values were rounded up to the nearest whole-Ampere and tenfold cross validation was used on the entire one year of data. The whole-house current draw measurement was denoised so that it equalled the summation of the current draw from the 11 loads chosen for disaggregation (a %-NM of 0.00). The classification and estimation results are listed in Table 2. We have also provided other basic measures in Table 1. Additionally, we include true-negatives tn counts, and the accurate/inaccurate true-positives (atp and itp) using in M-fscore, where $atp + itp = 1$ and can be seen as assigning partial accuracy and avoiding the binary nature of the true-positive tp score.

In all cases, basic accuracy scores far better than FS-fscore. This is most noted for the garage results. The $inacc$ results show partial penalization, and this is apparent when comparing f-score with FS-fscore. When we examine M-fscore, we see that it scores less than either f-score and FS-fscore, but it is hard to understand why. When examining the RMSE scores, it is hard to compare how appliances performed to each other or to the overall results as this score is not normalized. When comparing NDE with estimation accuracy, we see in most instances NDE scores better. This is most apparent in the ent/tv/dvd load. Overall, the FS-fscore and estimation of our test scores high, but this masks the fact some loads (clothes washer, garage, and home office) did not score well. Furthermore, the home office and garage results shows there can be a higher score for estimation but a lower classification score, and the ent/tv/dvd results show there can be a higher score for classification and a lower score for estimation.

Conclusion

We presented a unified approach that allows for consistent accuracy testing amongst NILM researchers. Our approach takes into account the classification performance and estimation performance—not one or the other. Additionally, we include performance reporting at both the overall level and an appliance level. This evaluation strategy has been incorporated into our research, and we look forward to continue the discussion and refinement of this framework as other NILM researchers continue to address the issue of inconsistent accuracy reporting.

Acknowledgments Research partly supported by grants from the Natural Sciences and Engineering Research Council (NSERC) of Canada, and the Graphics, Animation, and New Media Network of Centres of Excellence (GRAND NCE) of Canada.

References

Berges, M.E., Goldman, E., Matthews, H.S., Soibelman, L. (2010). Enhancing electricity audits in residential buildings with nonintrusive load monitoring. *Journal of Industrial Ecology*, *14*(5), 844–858.

Chang, H.H., Lin, C.L., Lee, J.K. (2010). Load identification in nonintrusive load monitoring using steady-state and turn-on transient energy algorithms. In *2010 14th International Conference on Computer Supported Cooperative Work in Design (CSCWD)* (pp. 27–32).

Dong, H., Wang, B., Lu, C.T. (2013). Deep sparse coding based recursive disaggregation model for water conservation. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence* (pp. 2804–2810): AAAI Press.

Figueiredo, M., de Almeida, A., Ribeiro, B. (2012). Home electrical signal disaggregation for non-intrusive load monitoring (nilm) systems. *Neurocomputing*, *96*(0), 66–73.

Johnson, M.J., & Willsky, A.S. (2013). Bayesian nonparametric hidden semi-markov models. *The Journal of Machine Learning Research*, *14*(1), 673–701.

Kim, H., Marwah, M., Arlitt, M., Lyon, G., Han, J. (2010). Unsupervised disaggregation of low frequency power measurements. In *11th International Conference on Data Mining* (pp. 747–758).

Kolter, J., & Johnson, M. (2011). Redd: A public data set for energy disaggregation research. In *Workshop on Data Mining Applications in Sustainability (SIGKDD)*. San Diego, CA.

Kolter, J.Z., & Jaakkola, T. (2012). Approximate inference in additive factorial hmms with application to energy disaggregation. In *International Conference on Artificial Intelligence and Statistics* (pp. 1472–1482).

Liu, H., & Motoda, H. (1998). *Feature selection for knowledge discovery and data mining*: Springer.

Makonin, S. (2014). Real-time embedded low-frequency load disaggregation. Ph.D. thesis, Simon Fraser University, School of Computing Science.

Makonin, S., Bajic, I.V., Popowich, F. (2014). Efficient Sparse Matrix Processing for Nonintrusive Load Monitoring (NILM). In *2nd International Workshop on Non-Intrusive Load Monitoring*.

Makonin, S., Popowich, F., Bartram, L., Gill, B., Bajic, I.V. (2013). AMPds: a public dataset for load disaggregation and eco-feedback research. In *2013 IEEE Electrical Power and Energy Conference (EPEC)* (pp. 1–6).

Parson, O., Ghosh, S., Weal, M., Rogers, A. (2012). Non-intrusive load monitoring using prior models of general appliance types. In *AAAI Conference on Artificial Intelligence*.

Tsai, M.S., & Lin, Y.H. (2012). Modern development of an adaptive non-intrusive appliance load monitoring system in electricity energy conservation. *Applied Energy*, *96*(0), 55–73.

Zeifman, M., & Roth, K. (2011). Nonintrusive appliance load monitoring: Review and outlook. *IEEE Transactions on Consumer Electronics*, *57*(1), 76–84.