Towards a new evolutionary subsampling technique for heuristic optimisation of load disaggregators

Michael Mayo and Sara Omranian

Dept. of Computer Science, University of Waikato
Hamilton, New Zealand
mmayo@waikato.ac.nz
sara.omranian@gmail.com

Abstract. In this paper we present some preliminary work towards the development of a new evolutionary subsampling technique for solving the non-intrusive load monitoring (NILM) problem. The NILM problem concerns using predictive algorithms to analyse whole-house energy usage measurements, so that individual appliance energy usages can be disaggregated. The motivation is to educate home owners about their energy usage. However, by their very nature, the datasets used in this research are are massively imbalanced in their target value distributions. Consequently standard machine learning techniques, which often rely on optimising for root mean squared error (RMSE), typically fail. We therefore propose the target-weighted RMSE (TW-RMSE) metric as an alternative fitness function for optimising load disaggregators, and show in a simple initial study in which random search is utilised that TW-RMSE is a metric that can be optimised, and therefore has the potential to be included in a larger evolutionary subsampling-based solution to this problem.

Keywords: non-intrusive load monitoring, disaggregation, imbalanced regression, fitness function, evolutionary undersampling

1 Introduction

A significant problem facing modern society is ensuring the efficient use of energy resources. One area where energy efficiency improvements can be made is in the domestic arena. If householders understand their own individual energy usages, down to the level of detail of individual appliances, then they are more likely to change their behaviour in way that conserves energy [11]. However, the main problem with this is that a household typically has tens of appliances and power outlets, and for accurate monitoring and reporting to householders, sub-meters must be physically installed at each outlet.

Non-intrusive load monitoring (NILM) is an alternative to this that attempts to replace sub-meters with machine learning algorithms. The algorithms are used

to train models that predict individual appliance/power outlet energy usage from the stream of whole-house energy usage that is provided by a typical smart meter.

Although this field has been a focus of research for many years (ever since Hart's seminal 1992 paper [5]), it has developed dramatically with the much more recent advent of smart meters capable of measuring household energy consumption at high frequency. New and massive datasets have been captured and are currently being analysed in research contexts.

To put the disaggregation problem into a more formal context, let us assume that there are n appliances and/or power outlets (hereafter simply referred to as an appliance), labelled $A^{(1)}, A^{(2)}, \ldots A^{(n)}$ and that ith appliance's energy usage at time t is $A_t^{(i)}$. The values of $A_t^{(i)}$ are unknown (except in the training data) and to be predicted for each appliance. The only values that are known are the stream of whole house energy use measurements $Z_1, Z_2, \ldots Z_{t-2}, Z_{t-1}, Z_t$ leading up to the current time (t).

We note that the unit of measurement can be any acceptable measure of energy usage. In this work, we adopt *current*, measured in amperes (A), as the stream to measure and predict. This follows Makonin et al.'s [7] analysis showing that current is the most reliable measure for data analysis because it fluctuates to a much lesser degree than other measures such as real power (which is frequently used in other research works in this field).

The remainder of this paper is organised as follows. In Section 2 a brief background provided. In Section 3 a new metric named target-weighted RMSE (TW-RMSE) is introduced to address the problem of imbalance that occurs in the of regression target variables in most disaggregation datasets. Then, in Section 4, a random search–based algorithm is presented to select a regression model that optimizes TW-RMSE. Section 5 evaluates the presented approaches on the AMPds dataset [7] and makes a complete comparison between RMSE and TW-RMSE. Lastly, Section 6 concludes the paper.

2 Background

2.1 Load disaggregation

Non-intrusive load monitoring/disaggregation research goes back many years. One of the first researchers who proposed a disaggregation method was Hart [5], who stated that in order to decompose the total load into its components, models of individual appliances and their combinations were needed. He was also the first to observe that the power consumed by appliances was additive, i.e. if the loads of all appliances in a household are known, then at any time t the following equality holds: $Z_t = \sum_i A_t^{(i)}$.

With the introduction of more powerful computers and significantly more memory and processing power, disaggregation research took a large step forward. Zeifman [12], for example, proposed an algorithm that uses stepwise power changes, power surges, and time-on and time-off durations as features. His approach used historical data for initial training of a disaggregation model. The

main aspect of his algorithm was the use of approximate semi-Markov models which could make robust computationally-efficient predictions. The downside of his algorithm was that it only worked with on/off appliances and not general appliances which may be in multiple discrete states, or transition continuously between states.

2.2 AmpDS v2 dataset

The Almanac of Minutely Power dataset ("AMPds") was introduced by Makonin et al. [7] in 2013, and a revision (version 2) released sometime afterwards. This dataset is a record of one house containing both total and submetered power data from 25 submeters at one minute intervals for two years (from 1 April 2012 until 31 March 2014). The dataset comprises 1,051,200 records and includes, besides the meter readings, additional data about water usage and climate. As Figure 1 depicts, the measurements of current draw which are the focus of our predictive experiments are highly imbalanced. The figure shows the log frequencies of current draws for a washing machine for one year's worth of data.

As mentioned in the Introduction, a major finding of the authors of the AMPds dataset and its associated paper was the fact that current provides better quality predictions than real power. Another new idea in this paper was the introduction of a novel method for discretising current draws, to fit the notion of appliances being like finite state machines (i.e. many appliances operate in discrete states, such as ON/OFF or LOW/MEDIUM/HIGH).

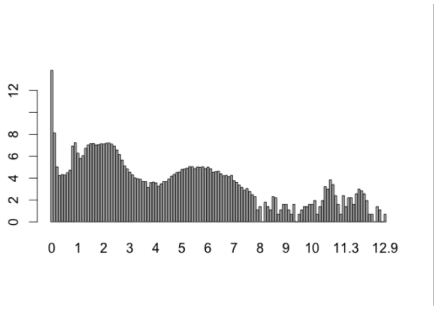

Fig. 1: Log frequencies of washing machine current draws. Shown are current draw (x) vs. the log frequency (in seconds) when that level of current usage was metered (y).

While a classification approach to solving this problem is reasonable, we focus in this paper instead on regression and attempt to predict appliance current draw directly. If a good regression-based solution to the problem is found, the approach can be adapted to the classification setting by discretising current using the approach described by Makonin et al. [7]. Another reason for adopting a regression-based approach to the problem is that the finite state machine abstraction is not true for all appliances: some devices clearly vary continuously in their energy usage, and even for appliances that do appear to operate in discrete states, the number of and boundaries between each state are not always clear (as Figure 1 attests).

2.3 Evolutionary machine learning for massive imbalanced regression problems

To date, few prior works exist on the problem of regression using imbalanced datasets [1]. This is because most of the classical imbalanced dataset research in machine learning concerns classification rather than regression. Furthermore, the authors are unaware of any prior evolutionary machine learning works on the imbalanced regression problem, let along any works considering the problem in the context of load disaggregation. This is a potentially rewarding area for research, therefore. Given the paucity of evolutionary approaches to this problem, some of the few non-evolutionary approaches will be briefly mentioned.

The earliest work on regression with imbalanced data is the paper by Torgo et al. [10] in which the standard SMOTE algorithm [3] was adapted to a regression setting. Essentially, the modified SMOTE algorithm under-sampled examples with target values that occur more frequently in the data. They also considered the two different ways of creating a balanced sample: under-sampling and over-sampling. They believed over-sampling was generally better than under-sampling, especially when small samples were involved. This made sense to the authors, since by using under-sampling full use of the available data was not made.

Beyond that, further works concerning imbalanced regression have only been published very recently. A common approach is to define new metrics for the problem As justification, Branco et al. [1] argue that performance measures commonly used in regression, such as Mean Squared Error (MSE) and Mean Absolute Deviation (MAD) are not adequate for imbalanced regression problems. This is because they assume a uniform relevance of the target variable domain and evaluate only the magnitude (and not the direction) of the error.

One approach was presented by Hernández-Orallo [6], which introduced ROC space for regression (RROC). RROC space is defined by plotting total over-estimation and under-estimation on the x and y axes respectively. The author also proposed area over the RROC curve (AOC) as a metric, but the drawback of his approach was it only accounted for under-predictions. Another metric proposed by Ribeiro [9] and Torgo and Ribeiro [10] is based on the concept of utility-based regression for precision/recall metrics. Utility-based regression uses the notion of non-uniform relevance of the target variable values across the domain of the target to define a metric. For example, certain parts of the target variable domain may be more critical than others; therefore the utility is higher. This notion of utility led to the proposal of metrics such as the Mean Utility and the Normalized Mean Utility by Riberio [9], and then to a utility-based regression algorithm based on regression rule ensembles called ubaRules [9].

3 Target-weighted RMSE metric

To address the problem of massive imbalance that occurs in the AMPds dataset we propose a modification to the standard RMSE metric for evaluating regressors. We name the new metric target-weighted RMSE (TW-RMSE).

The basic idea is related to the notion of macroaveraged accuracy [8] in information retrieval. Macroaveraged accuracy is a alternative metric for evaluating classification models in which individual classes are weighted equally in the accuracy calculation, as opposed to the weighting being on individual examples (which is the case for normal or "micro"-averaged accuracy). As Manning et al. state, common classification metrics such as the F1 measure fail when the imbalance is overwhelming. And this is largely because the F1 metric is a micro-averaging approach. Macroaveraging, on the other hand, is an attempt to eliminate the problem of massive class imbalance.

Since our problem domain is regression, however, we do not have readily available class labels to make use of. Therefore we define proxy class labels for each example by dividing the range of the predictive target variable into equal-width bins, and we treat an example's bin as its class label for the purposes of balancing, resampling, and computing performance metrics.

Furthermore, we allow the user to specify (i) the number of bins and (ii) an individual weight for each bin. The weights serve a similar purpose to utility in utility-based regression, in that they allow the user to focus the metric on particular ranges of current draw. For example, if the user is only interested in particular states that an appliance may be in, then the weights for the corresponding ranges of current usage can be increased when the metric is calculated.

Once the ranges and weights are set, an RMSE value is calculated for test example predictions in each range. The TW-RMSE is the weighted average of these individual RMSE values. Algorithm 1 shows pseudocode for performing this calculation.

Input: $\mathbf{Y} = \{(y_1, y_1^*), (y_2, y_2^*) \dots (y_n, y_n^*)\}$, a vector of predicted (y_i) and actual (y_i^*) values for n training or test examples; m, number of bins; $\mathbf{w} = \{w_1, w_2 \dots w_m\}$, a vector of weights for the bins

begin

$l \leftarrow min(y_1^*, y_2^* \dots y_n^*)$;

$u \leftarrow max(y_1^*, y_2^* \dots y_n^*)$;

Divide the range $(u - l)$ into m equal-width bins $\mathbf{B} = \{b_1, b_2 \dots b_m\}$;

Assign each $(y_i, y_i^*) \in \mathbf{Y}$ to the appropriate bin in $\mathbf{B}$ (using the y_i^* value);

Delete any empty bins from $\mathbf{B}$;

Compute the RMSE $r(b_j)$ for the predictions in each $b_j \in \mathbf{B}$;

$t \leftarrow \sum_{b_j \in \mathbf{B}} w_j r(b_j)$ where $w_j \in \mathbf{w}$ is bin b_j's weight;

end

Output: t, the TW-RMSE metric

Algorithm 1: Algorithm to compute the TW-RMSE metric.

Effectively, this approach should give an unbiased error estimate if the dataset is imbalanced. As the number of bins is reduced, however, the TW-RMSE metric becomes an approximation of the RMSE metric: therefore the user must ensure that the number of bins is set correctly.

4 Random search-based algorithm to optimise TW-RMSE

To demonstrate the potential utility of the TW-RMSE metric for appliance energy usage modelling, we propose a simple random resampling-based algorithm that uses TW-RMSE to select an optimal training sample for a regressor.

The approach can be divided into two parts. Firstly, Algorithm 2 depicts a method for creating a balanced resample from the original, highly imbalanced, training dataset. The approach taken is similar to the algorithm for computing TW-RMSE, except that instead of binning predictions, this time the examples in the training set are iteratively resampled into bins that are the same size. A parameter s governs the overall size of the sample.

Input: $\mathbf{X} = \{(\mathbf{x_1}, y_1^*), (\mathbf{x_2}, y_2^*) \dots (\mathbf{x_n}, y_n^*)\}$, a training dataset of examples $(\mathbf{x_i})$ and targets (y_i^*); m, number of bins; s, desired sample size ($s << n$);
begin
  $l \leftarrow min(y_1^*, y_2^* \dots y_n^*)$;
  $u \leftarrow max(y_1^*, y_2^* \dots y_n^*)$;
  Divide the range $(u - l)$ into m equal-width bins $\mathbf{B} = \{b_1, b_2 \dots b_m\}$;
  Assign each example $(\mathbf{x_i}, y_i^*) \in \mathbf{X}$ to the appropriate bin in $\mathbf{B}$ (using the y_i^* value);
  Delete any empty bins from $\mathbf{B}$;
  $\mathbf{S} \leftarrow \emptyset$;
  repeat
    foreach $b \in \mathbf{B}$ **do**
      if $|\mathbf{S}| < s$ **then**
        Randomly resample one example from b into $\mathbf{S}$
      end
    end
  until $|\mathbf{S}| \geq s$;
end
Output: $\mathbf{S}$, a balanced resampling of the dataset

Algorithm 2: Algorithm to create a balanced resample of examples.

The second part of the algorithm is a simple random search procedure, pseudocode of which Algorithm 3 depicts. The algorithm essentially repeatedly resamples the entire training dataset, creating a small balanced resample with each iteration. A regressor is then trained on each sample, and the TW-RMSE metric computed for the sample. The regressor with the smallest TW-RMSE is returned after a fixed number of iterations.

In the next section, it will be shown that this approach results in a significant decrease in TW-RMSE for most appliances, but that RMSE on the whole is unchanged (thus validating our rationale for TW-RMSE).

Input: $\mathbf{X} = \{(\mathbf{x_1}, y_1^*), (\mathbf{x_2}, y_2^*) \dots (\mathbf{x_n}, y_n^*)\}$, a training dataset of examples ($\mathbf{x_i}$) and targets (y_i^*); m, number of bins; s, desired sample size ($s << n$); i, number of iterations

begin

  $\mathbf{S}_{best} \leftarrow$ a subset of the training data $\mathbf{X}$ selected using Algorithm 2;

  $R \leftarrow$ a regressor trained on $\mathbf{S}_{best}$;

  $t_{best} \leftarrow$ TW-RMSE of R_{best} evaluated against $\mathbf{S}_{best}$ calculated using Algorithm 1;

  for *iteration* $\leftarrow 2 \dots i$ **do**

    $\mathbf{S} \leftarrow$ a subset of the training data $\mathbf{X}$ selected using Algorithm 2;

    $R \leftarrow$ a regressor trained on $\mathbf{S}$;

    $t \leftarrow$ TW-RMSE of R evaluated against $\mathbf{S}$ calculated using Algorithm 1;

    if $t < t_{best}$ **then**

      $R_{best} \leftarrow R$

    end

  end

end

Output: R_{best}, the best regressor obtained.

Algorithm 3: Training algorithm based on random search and model selection.

5 Evaluation

In order to evaluate the effectiveness of TW-RMSE as a metric for evaluating and selecting regression models in the presence of imbalance, we performed an extensive set of experiments on portions of the AMPds v2 dataset [7]. The dataset itself contains energy usage data (including current, real power and voltage) for a large number of sub-metered appliances.

We selected nine sub-meters to analyse, specifically the same sub-meters as previously analysed in a classification setting by Makonin et. al. [7]: the basement plugs and lights (BME), the clothes dryer (CDE), the clothes washer (CWE), the dishwasher (DWE), the kitchen fridge (FGE), the HVAC/furnace (FRE), the heat pump (HPE), the entertainment unit (TVE) and finally the wall oven (WOE). In each case, we extracted and built predictive models for the current (A) time series only. These formed our (unknown) $A_t^{(i)}$ values for the evaluation. We also extracted from the dataset the whole-house (WHE) current draw, which were our (known) Z_t values. As mentioned previously, the total length of each time series, covering two years worth of data sampled once per second in each case, is 1,051,201 samples.

A sliding window approach was used to generate a dataset suitable for learning regression models. Essentially, this approach generates examples by "sliding" a fixed-size window along the WHE series, creating one each example for each position that the window can occupy. Since the window size we chose was fixed at 60, each example's features therefore correspond to whole-house energy usage for the past 60 seconds.

Unlike standard time series prediction, however, the prediction target for each example is not the next WHE value. Instead it is the corresponding sub-meter reading. Notation-wise, the features are $Z_{t-59}, Z_{t-58} \ldots Z_{t-1}, Z_t$ and the prediction target is $A_t^{(i)}$ where i is the sub-meter index and t is the current time in seconds.

This procedure produced a dataset with (1,051,201 - 60 =) 1,051,141 examples. Since the examples are ordered in time (and therefore the i.i.d. assumption does not hold), we further divided these examples into two subsequences: the first 50% of examples (a year's worth of data) were used for training, and the second 50% (a second year's worth of data) for testing. A total of nine datasets for the nine sub-metered appliances were created in this way.

Fig 2 shows distributions of current in the test data by appliance, which reinforces seriously imbalanced nature of problem.

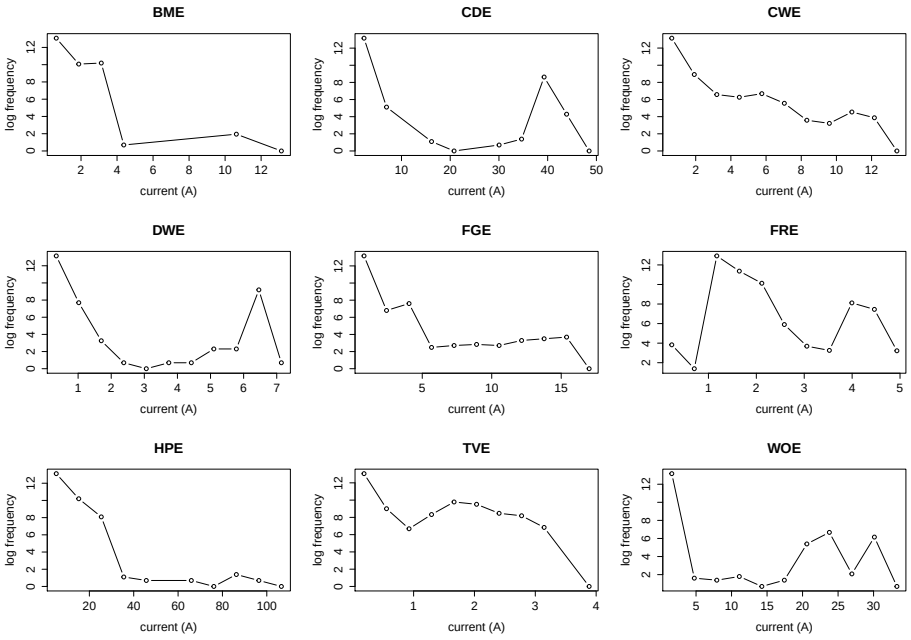

Fig. 2: Distribution of target (i.e. appliance current) values in the test sets. Note that the y axis shows the *log* frequencies rather than the raw frequencies.

We used the standard random forest classifier adapted for regression for our experiments [2]. The number of random trees in the forest is set to 100, and the number of random features selected per tree was $\lfloor log_2(60) + 1 \rfloor = 7$.

In our first experiment, we trained a random forest regressor on each *entire* training set (i.e. we ignored Algorithms 2 and 3), and then we evaluated its performance on the corresponding test sets.

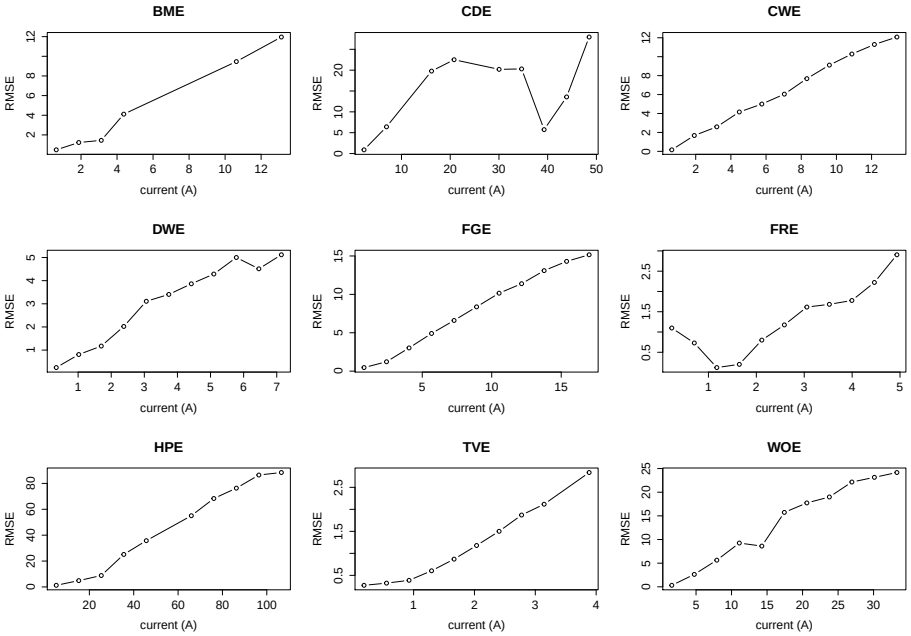

Fig. 3: RMSE values calculated for each bin in the test sets, using regressor trained on full training sets.

Appliance	RMSE	TW-RMSE
BME	0.62	4.78
CDE	1.09	15.26
CWE	0.43	6.37
DWE	0.66	3.05
FGE	0.54	8.06
FRE	0.29	1.30
HPE	1.82	45.04
TVE	0.43	1.20
WOE	1.13	13.49

Table 1: Test results after training a random forest classifier on the entire training dataset.

Figure 3 shows the distribution of RMSE by bin across the test data for this first experiment. This figure illustrates nicely how error generally increases with current draw. Increased current draw usually corresponds to smaller frequencies in the datasets.

Table 1 gives values for both RMSE and TW-RMSE (using $m = 10$ bins and uniform $w_i = \frac{1}{m}$ weights) computed on the test datasets. These values are useful for determining the effects of Algorithms 2 and 3 in our next experiments.

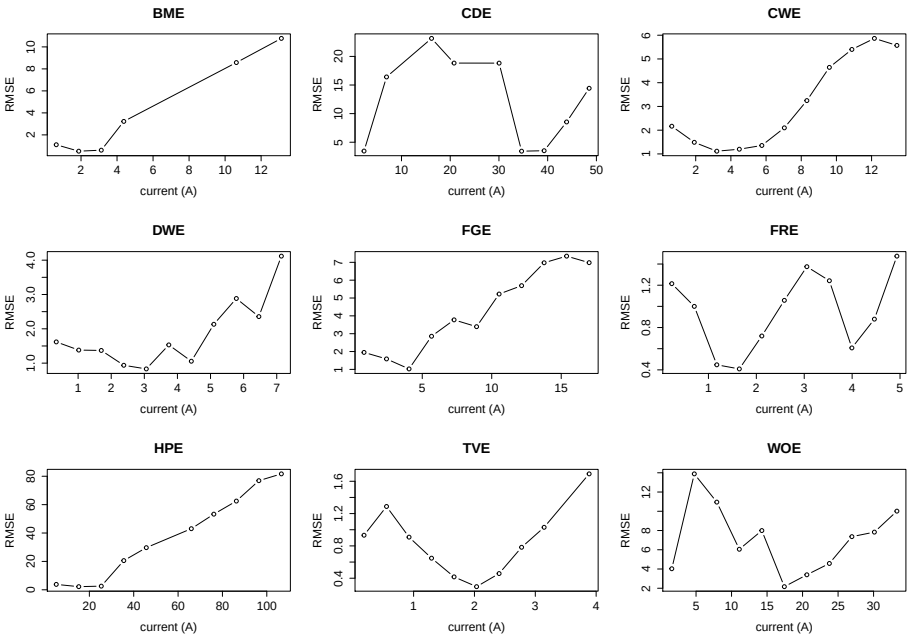

Fig. 4: RMSE values calculated for each bin in the test sets, using regressor trained on a balanced resample of the training set selected using Algorithm 2.

Next, we trained the random forests model on a single resample of the training data (obtained using Algorithm 2) and tested it against the entire test dataset. The number of bins m and the weights w_i are the same as in the previous experiment, and we set the sample size $s = 10,000$. Figure 4 and Table 2 depict the results.

Two observations can be made from these results. Firstly, TW-RMSE is reduced by a significant amount when the regressor is trained on the small resampled training dataset compared to being trained on the entire training dataset. This is clearly evident graphically when Figures 3 and 4 are compared: for many appliances, e.g. the wall oven (WOE), the error curve no long approximately increases with increasing current draw. This graphical difference is reflected in

Appliance	RMSE	TW-RMSE
BME	1.07	$4.13^{\dagger\dagger}$
CDE	3.47	$12.29^{\dagger\dagger}$
CWE	2.16	$3.10^{\dagger\dagger}$
DWE	1.63	$1.84^{\dagger\dagger}$
FGE	1.95	$4.26^{\dagger\dagger}$
FRE	0.46	$0.95^{\dagger\dagger}$
HPE	3.74	$37.63^{\dagger\dagger}$
TVE	0.91	$0.85^{\dagger\dagger}$
WOE	4.03	$7.11^{\dagger\dagger}$

Table 2: Test results after training a random forest classifier on a balanced resample of the training dataset selected using Algorithm 2. Results marked † indicate an improvement compared to training on the entire training dataset. Results marked †† indicate that the improvement is more than 10%.

changes in the TW-RMSE metric which in all cases is more that 10% and sometimes 50% reduced in magnitude (comparing Tables 1 and 2).

In contrast, comparison between the tables also shows that the RMSE metric does not decrease at all. In fact, the more balanced subsampling results in an overall increase of RMSE because of the much greater emphasis given to the majority of examples in which the appliance is in an off or low energy state.

Appliance	RMSE	TW-RMSE
BME	1.07	$3.40^{\dagger\dagger}$
CDE	$3.34^{\dagger}$	$12.02^{\dagger}$
CWE	2.18	3.13
DWE	$1.57^{\dagger}$	$1.78^{\dagger}$
FGE	$1.94^{\dagger}$	4.27
FRE	0.46	$0.67^{\dagger\dagger}$
HPE	3.84	$32.08^{\dagger\dagger}$
TVE	$0.90^{\dagger}$	$0.45^{\dagger\dagger}$
WOE	4.31	7.24

Table 3: Test results after training a random forest classifier using Algorithm 3. Results marked † indicate an improvement compared to training on only a single balanced sample. Results marked †† indicate that the improvement is more than 10%.

Finally, we ran Algorithm 3 for $i = 100$ iterations, and the results are shown in Table 3. Comparing to Table 2, there is a further marked decrease in TW-RMSE for seven of the sub-meters as a consequence of the random search. The only appliances that did not show improvement when Algorithm 3 was applied with 100 as opposed to 1 iteration were the clothes washer (CWE), the kitchen fridge (FGE) and the wall oven (WOE). However, on the whole, most of the sub-meter predictions further improve compared to Table 2. This show that the

TW-RMSE metric, when computed on a training set, is a good indicator of generalisation performance.

6 Conclusion

This work represents our first step towards the development of an evolutionary subsampling technique for solving the NILM problem. The main contribution of interest here is a new fitness metric, namely TW-RMSE, which can counter the effect of massive imbalance that occurs in typical NILM datasets. A further lesser contribution is an algorithm inspired by TW-RMSE for generating balanced random under-samples of examples.

The results here are promising and demonstrate that TW-RMSE has potential to be a useful fitness metric. Our next steps will be (i) to investigate the potential sensitivity of the metric to its key parameters, such as the number of bins; and (ii) to couple TW-RMSE with a more state-of-the-art evolutionary subsampling algorithm (e.g. a method reviewed by Derrac et al. [4]) in order to more properly evaluate its potential.

References

1. Branco, P., Torgo, L., Ribeiro, R.: A survey of predictive modelling under imbalanced distributions. arXiv:1505.01658v2 (2015)
2. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
3. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: Synthetic minority over-sampling technique. J. Artif. Int. Res. 16(1), 321–357 (2002)
4. Derrac, J., Garcia, S., Herrera, F.: A survey of evolutionary instance selection and generation. International Journal of Applied Metaheuristic Computing 1(1), 60–92 (2010)
5. Hart, G.W.: Nonintrusive appliance load monitoring. Proceedings of the IEEE 80(12), 1870–1891 (1992)
6. Hernández-Orallo, J.: Roc curves for regression. Pattern Recognition 46(12), 3395–3411 (2013)
7. Makonin, S., Popowich, F., Bartram, L., Gill, B., Bajic, I.: AMPds: A public dataset for load disaggregation and eco-feedback research. In: Electrical Power Energy Conference (EPEC), 2013 IEEE. pp. 1–6 (2013)
8. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press (2008)
9. Ribeiro, R.: Utility-based regression. Ph.D. thesis, Dep. Computer Science, Faculty of Sciences, University of Porto (2011)
10. Torgo, L., Ribeiro, R.: Utility-based regression. In: Knowledge Discovery in Databases: PKDD 2007. pp. 597–604. Springer Berlin Heidelberg (2007)
11. Vine, D., Buys, L., Morris, P.: The effectiveness of energy feedback for conservation and peak demand: A literature review. Open Journal of Energy Efficiency 2, 7–15 (2013)
12. Zeifman, M., Roth, K.: Nonintrusive appliance load monitoring: Review and outlook. Consumer Electronics, IEEE Transactions on 57, 76–84 (2011)