Noname manuscript No.
(will be inserted by the editor)

On time series representations for multi-label NILM

Christoforos Nalmpantis · Dimitris Vrakas

Received: date / Accepted: date

Abstract Given only the main power consumption of a household, a NILM system identifies which appliances are operating. With the rise of Internet of Things, running energy disaggregation models on the edge, is more and more mandatory for privacy concerns and economical reasons. However, current NILM solutions use data hungry Deep Learning models that can recognize only one device. This research investigates in depth multi-label NILM systems and suggests a novel methodology for a more cost effective solution. The key to this proposal is the introduction of popular time series representations, among them the novel Signal2Vec framework. Experiments are conducted to compare these representations and the two best ones, Signal2Vec and BOSS, perform on paar with previous multi-label NILM systems.

Keywords nonintrusive load monitoring · multi-label NILM · smart grids · time-series representation · signal2vec

This research has been co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH–CREATE–INNOVATE (project code:95699 - Energy Controlling Voice Enabled Intelligent Smart Home Ecosystem).

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

C. Nalmantis
School of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece
E-mail: christofn@csd.auth.gr

D. Vrakas
School of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece
E-mail: dvrakas@csd.auth.gr

1 Introduction

Efficient and accurate energy monitoring of domestic buildings can significantly diminish energy wastage. In the US the savings of the electricity consumption by residential buildings is estimated around 15 % of the total consumption or 200 billion kWh per annum. This amount is equivalent to 81.3 million tons of coal [12]. The advantages of energy monitoring are not restricted to the positive effect on the environment. It will affect many sectors in the industry by reforming building operations, advancing smart grid optimization, boosting energy consumption forecasting etc. [2].

Non-intrusive load monitoring has been a subject of research since 1992 [14]. It has been favored over intrusive methods due to economic and practical reasons. In this context, a NILM system needs to meet the requirements of a large-scale deployment in the real world. As far as hardware is concerned, most NILM systems are applicable because they offer lower costs and simpler installation. On the software side though, power disaggregation models can be computationally intensive and thus not economically viable. This is not a surprise as power disaggregation has been classified as NP-complete [14].

Numerous NILM approaches have been proposed ranging from combinatorial optimisation techniques [3, 25, 44] to probabilistic models [53, 29, 52], machine learning methods [37, 43, 15, 21] and other methodologies [10, 20, 25]. Most recent approaches include variations of Hidden Markov Models and deep neural networks. HMMs were the state-of-the-art for long time but their main drawback is that they are computationally intractable and do not scale for large number of appliances. Deep neural networks have only recently been proposed and have demonstrated very promising results [31]. The most popular deep learning architectures are recurrent neural networks (RNNs) [16] and convolutional neural networks (CNNs) [50]. Although comparing different NILM solutions has been a non trivial task [31], recent benchmarks have shown that deep learning methods clearly demonstrate state-of-the-art performance [9].

Despite the high accuracy results, the deployment of a NILM system based on deep neural networks would require huge computing power. The root cause is manifold. Firstly, deep neural nets consist of millions of parameters, which makes the processes of learning and inference heavy. Secondly, energy data are generated every second or minute, contributing to a vast amount of data. Thirdly, according to the majority of the proposed solutions, identifying an appliance requires a model dedicated to this appliance. Consequently, even a distilled version of a deep neural network, like the one named "online GRU" [19], would require a one-to-one model-appliance relationship.

The aforementioned problems make running NILM models at edge prohibitive, whereas a cloud solution would be very expensive. A viable solution should have the following properties: a) A lightweight disaggregation model, b) A data representation which achieves sufficient dimensionality reduction, c) A one-to-many relationship, where one model can identify many appliances. The latter property can be addressed by handling NILM as a multi-label classification problem. According to the literature, very few similar approaches have been proposed so far. The other two properties can be met by evaluating various lightweight multi-label machine learning models in combination with time series approximation methods.

In this research, seven of the most popular time series representations in the domain of time series classification are selected. To the best of the authors' knowledge, this is the first time these algorithms are used in the problem of multi-label NILM. In addition, Signal2Vec, a novel framework proposed in our previous work [32], is compared and demonstrates the most promising results.

The main contributions of this paper are: a) Literature review of the existing systems facing NILM as a multi-label classification problem. b)Analysis of the novel framework Signal2vec, c) Detailed evaluation of eight different time series representations with respect to the problem of multi-label NILM.

The remainder of this article is organized as follows. Firstly the review of related work is presented with a focus on multi-label solutions. Next, there is a short description of the seven existing time series representations and a detailed analysis of the novel framework Signal2Vec. In the last section, the conducted experiments are described with details around the methodology, an in depth analysis of the results and comparison with other solutions. Finally there is a summary and suggestions for future work.

2 Multi-label NILM

The goal of a NILM system is to break the total power consumption of a residence, into estimates of the actual power demand of each appliance. It is a single source separation problem and every sample is a mixture of signals, that are not mutually exclusive. Considering now each target appliance as a binary label, the problem of power disaggregation can be seen as a multi-label classification task.

Multi-label classification was firstly introduced in NILM by Basu et al. [4, 5]. The proposed method used the aggregate power consumption at 10 minutes interval and identified three energy-hungry appliances. The algorithms under examination were: decision trees [47], boosting [49], Rakel [47], MLkNN [51] and classifier chains [34]. During testing phase the models showed encouraging results as far as the test period was no more than 6 months. For longer period the performance dropped significantly and was attributed to the seasonal variations of pattern usage of appliances. Thus, a model in the real world would be retrained at least every six months. A similar approach was used in [6, 7], where multi-label classification models were compared against a standard single label HMM algorithm.

Li et al. [24] proposed a different approach for multi-label NILM using an expectation maximization semi supervised method and converting energy time series to delay embeddings. The usage of delay embedding via Takens' theorem [46] has been very successful and has been incorporated in more multi-label NILM systems [22]. Tabatabaei et al. [45] compared delay embedding and wavelet features in combination with RAkEL and MLkNN. The experiments disaggregated 6-7 appliances and MLkNN in time domain demonstrated the best macro F1 score.

Recently more advanced machine learning methods pushed the limits of multi-label NILM with new state-of-the-art results. Li et al. [23] suggested three new graph-based semi-supervised multi-label algorithms for the problem of power disaggregation. According to the authors the suggested algorithms require quadratic time and even down sampling the data the duration of the experiments was very long. Finally, recent research employed deep learning methods and demonstrated very promising results, using novel multi-label methods such as RBM, deep dictionary learning and deep transform learning [48, 42, 41]

NILM has not been fully explored as a multi-label problem and there is plenty room of improvement. An important factor to consider in this research is the amount of data and the hardware that will run the disaggregation algorithms or the economical impact. Speaking of that, new proposals should focus on time series approximation methods and employ models with low computing demands.

3 Time series representations

In the era of Internet-of-Things there is an explosion of data. By 2020, it is expected that more than 50 billion IoT devices will be connected [35]. Handling the tremendous amount of information and extracting useful information using data mining techniques is necessary more than ever, but at the same time very challenging. One of the first steps to tackle this challenge of big data is to reduce the dimensionality. It is true that a large portion of this information is in the form of time series and there are various methods of time series dimensionality reduction. In the context of the problem of power disaggregation we propose the employment of a seven time series representations, which are briefly presented below.

Piecewise Aggregate Approximation (PAA) One of the most popular time series approximation is Piecewise Aggregate Approximation (PAA) [17]. The implementation is very straightforward. Firstly the series is divided into M time-frames of equal size. Then, the mean value of each time-frame is calculated, forming a representation of M dimensions. The algorithm of PAA is expressed as follows. Denote a time series $X = x_1, ..., x_n$, with length n. PAA is an approximation that represents time series X in M space by a vector $\bar{X} = (\bar{x}_1, ..., \bar{x}_M)$, where $M \leq n$. The ith element of $\bar{X}$ is calculated by the following equation:

$$\bar{x}_i = \frac{M}{n} \sum_{j=n/M(i-1)+1}^{(n/M)i} x_j \tag{1}$$

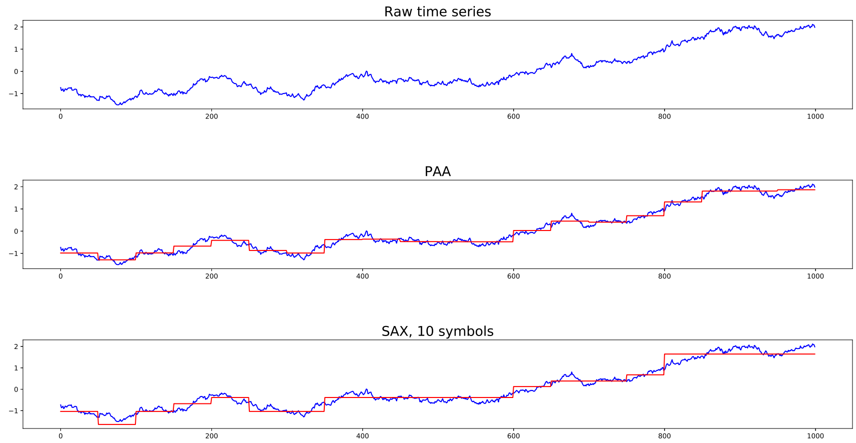


Fig. 1: Illustration of a time series and its PAA and SAX representations.

Symbolic Aggregate Approximation (SAX) SAX is an extension of PAA, inheriting its simplicity and the low computational complexity [26]. The main advantage of this algorithm is that it transforms a time series into a string representation. SAX has been very successful in many tasks related to time series such as classification, clustering, summarization, anomaly detection and indexing. The steps of the algorithms are: a) Apply PAA algorithm to segment the time series. b) Discretize the average values by mapping them to letters of an alphabet. The only requirement of the latter step is that the new symbols have to be produced with equiprobability. This is achieved with normalized time series because they have a Gaussian distribution. More formally this is expressed by defining breakpoints as a list of ordered numbers $B = \beta_1, \beta_2, ..., \beta_{a-1}$ such that the area under the Gaussian curve between two adjacent breakpoints is constant. Consequently, a time series can be converted into words by assigning a symbol $alpha_j$ for each interval $[\beta_{j-1}, \beta_j)$. The mapping of PAA approximation $\bar{C}$ into a string $\hat{C}$ is given by the formula below:

$$\hat{c} * i = alpha * j, \; iif \; \bar{c} * i \in [\beta_{j-1}, \beta_j) \tag{2}$$

1d-SAX An alternative version of SAX is 1d-SAX [28]. The benefit of this version is that it takes into account the trend of each subsequence when mapping the values of the time series into symbols. Thus after applying PAA, instead of computing the average of a segment, we compute the linear regression. The two coefficients of linear regression for each segment are mapped into symbols independently and then combined together into one symbol. The statistical properties of the average values and the slope values form a Gaussian distribution and consequently the quantization step is achievable in the same way as in the original version of SAX.

Discrete Fourier Transform (DFT) One of the most common time series approximation algorithms is the Discrete Fourier Transform (DFT), which was firstly used for the problem of time series similarity search by Agrawal et al. [1]. The most important properties of DFT are that its coefficients consist an orthogonal basis, the first few coefficients contain most of the information, it is fast to compute due to Fast Fourier Transform (FFT) algorithm and it preserves the Euclidean distance due to Parseval's theorem.

Symbolic Fourier Approximation (SFA) Schäfer et al. [39], inspired by DFT, suggested a novel time series representation named Symbolic Fourier Approximation. SFA consists of two steps, preprocessing and transformation. Preprocessing includes the DFT approximation and a quantization technique called multiple coefficient binning (MCB). MCB minimizes information loss during quantization by grouping the DFT coefficients of all subsequences, creating a histogram for each group and applying binning to each group. The transformation step maps the MCB discretization into symbols of a finite alphabet.

Bag-of-SFA-Symbols (BOSS) An extension and improvement of SFA is a time series representation named Bag-of-SFA-Symbols (BOSS), which is very robust in noise [38]. This feature is important in time series tasks, as real world data tend to be very noisy or erroneous. In order to obtain BOSS representation of a given time series T, sliding windows $S_{i;w}$ of size w are extracted. Next each window is converted to unordered SFA words. The transformations $SFA(S_{i;w}) \in \Sigma^l, fori = 1, 2, ..., (n - w + 1)$ are used to build a histogram. The BOSS histogram is a function, $B : \Sigma^l \to N$, which maps the SFA word space into the space of natural numbers. BOSS has the property of phase shift invariance because it doesn't take into account the order of SFA words and uses numerosity reduction [27,26] to avoid outweighing segments with constant values.

Word ExtrAction for time SEries cLassification (WEASEL) WEASEL is a novel time series representation which has demonstrated very promising results on the task of time series classification [40]. The efficiency of this method is attributed to two new methods named discriminative approximation and discriminative quantization. The former one uses the one-way ANOVA F-test during the approximation step, to select the Fourier coefficients that characteristic to class labels. The latter method maximizes the information gain, resulting in low entropy feature set of class labels. Given a time series, WEASEL firstly extracts normalized windows of various sizes. Then Fourier coefficients are calculated and filtered, keeping the ones that are characteristic to this particular time series. The filtering process is achieved by using the ANOVA F-test. Next, the Fourier coefficients are quantized. The quantization process is completed using information gain binning to best separate the various time series classes. The unigrams and bigrams are then used to construct a bag-of-patterns, filtering out irrelevant words using the Chi-Squared test.

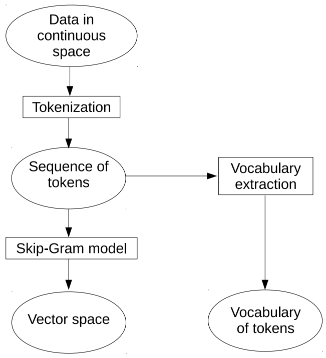


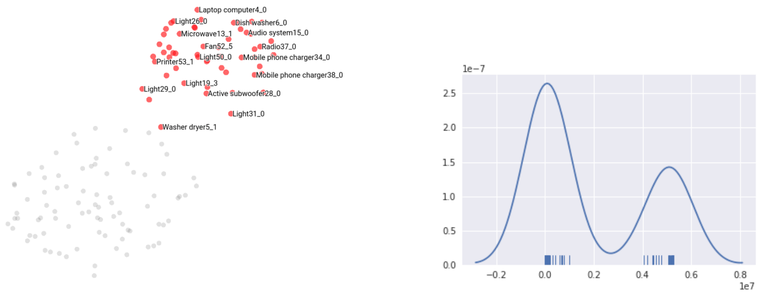
Fig. 2: Signal2Vec framework.

Bag-of-patterns approaches have linear computational complexity, therefore WEASEL achieves the best tradeoff between accuracy and speed.

4 Signal2Vec

Signal2vec [32] is a novel framework, that maps any time-series into a vector space. It is a generalization of word2vec [30] and extends its applicability to sequences in continuous space. As most of other time series representation algorithms, the first step of Signal2vec is a quantization process. Next, the skip-gram model [30] is applied to build a vector space. The framework is depicted in Figure 2. The learning process of the embedding feature space is unsupervised, computationally efficient and scalable. The main advantage in comparison to other time series representation algorithms is that the embeddings are built only once and can be reused in unseen datasets. Another important property is that the new constructed embedding encapsulates the frequency distribution of the quantized sequence and hence the frequency features of the original signal, Figure 3.

The given time series is quantized via clustering and each cluster defines a token. The number of clusters is estimated using silhouette score, which tells which objects fit well in their cluster [36]. The search space of the number of clusters is also constrained by a minimum and a maximum desirable number of clusters. In the domain of energy buildings the hypothetical bounds of the space can be estimated by calculating the number of possible energy states using the complexity of power draws [11]. Once the clusters have been found, a function is trained to map any time series to a discrete sequence of tokens. This function can be a classifier, like k-nearest neighbors, which is trained only



(a) Visualization of Signal2Vec embeddings using TSNE.

(b) Frequency distribution of tokens.

Fig. 3: Signal2Vec encapsulates the frequency distribution of the tokens.

once and then can be used to map any new time series to the finite space of tokens.

In the terminology of word2vec, the finite space of tokens is called vocabulary and a sequence of tokens is called corpus. In order to apply skip-gram model we need to define a context.

Definition 1 *Context:* Given a sequence of tokens and a sliding window of length W, with W odd, let TKN_{target} be the middle token of the window. Then, the context consists of the rest of the tokens inside the window.

The objective of the skip-gram model is to predict the context given a specific token. The architecture is a shallow neural network with one hidden layer and is trained with pairs of tokens. One token is the target and the other one belongs to the context. The network trains the weights of the hidden layer from the frequencies each pairing shows up. The formal definition of the objective function is given below.

Let $TKN_1, TKN_2, ...TKN_n$ be a sequence of N training tokens. Then the objective function tries to maximize the average log probability according to the formula:

$$1/N \sum_{t=1}^{N} \sum_{-c \leq i \leq c, i \neq 0} \log p(TKN_{t+i} | TKN_t) \tag{3}$$

, where c is the training context.

The loss function of the neural network is the Noise Contrastive Estimation (NCE) [13] and the optimizer is the Adagrad with learning rate 0.001. The size of each embedding is 300 and the context is 6 tokens.

Definition 2 *Signal2Vec:* Let a time series S of length N, a mapping function $f : R \rightarrow T^n$ and a look up function $g : T^n \rightarrow V^{n*M}$, with T a finite set of n tokens and V an $M - dimensional$ vector space with n vectors. We define Signal2Vec as the mapping of a time series into a vector space with the following formula:

$$Signal2Vec = g(f) : R \rightarrow V^{n*M} \tag{4}$$

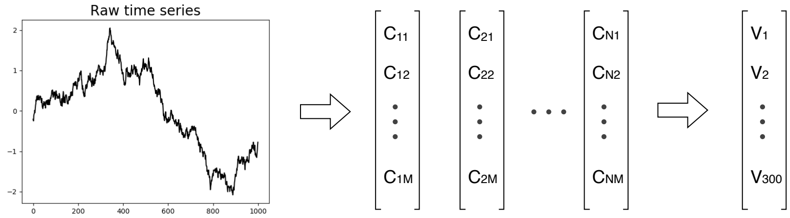


Fig. 4: Mapping a time series to one vector.

Following the definition of Signal2Vec, the constructed vector space has as many vectors as the size of the set of tokens, with each vector having 300 dimensions. Thus, a time series is mapped to a sequence of tokens, which is mapped into a sequence of 300-dimensional vectors. Such a representation is not very helpful as it worsens the problem of dimensionality explosion. The concept of the average vector representation solves this problem.

Definition 3 *Average vector representation:* Let a time series S of length N mapped into a sequence of N $M - dimensional$ vectors $\hat{C_i}, i = 1, ..., N$. The average vector representation $S2V_{avg}$ of the time series S is defined by:

$$S2V_{avg} = 1/N \sum_{i=1}^{N} C_i \tag{5}$$

Consequently a time series of length N can be represented by just one $M - dimensional$ vector, like in Figure 4. The benefit of this mapping is that the final representation is independent from the initial length N. In case a time series is very long, it can be broken into equal segments and the final vector representation will be the concatenation of the average vectors of each segment.

Training Signal2Vec is fast because the neural net is very shallow and it doesn't slow inference because it happens only once. Therefore, inference time mainly depends on the inference time of the classifier as the mapping of tokens to embeddings is $O(1)$. Thus, the inference time complexity if k-NN is $O(N * S)$, where N is the length of the time series under transformation and S is the number of samples the classifier was trained. Considering $S << N$, we conclude that time complexity of Signal2Vec inference process is linear.

5 Experiments and results.

The goal of the experiments is twofold. Firstly, to show that dimensionality reduction of energy data leads to computationally light solutions for the task of multi-label NILM. Secondly, to evaluate the effectiveness of eight different time series representations on this particular problem. The empirical results are also compared to others in the literature.

5.1 Data, metrics and methodology

The energy data that are used in the experiments come from two popular datasets UK-DALE [16] and REDD [18]. The latter one contains energy data from houses in the US and the former one from UK houses. For parsing the data we use NILMTK [8]. All experiments, including details of the implementation and results are available at https://github.com/ChristoferNal/available-after-acceptance.

The examined time series, as mentioned in previous sections are: PAA, SAX, 1d-SAX, DFT, SFA, BOSS, WEASEL and Signal2Vec. The only one that is suitable for transfer learning is Signal2Vec and for this reason we construct the embedding space only once. The vectors are produced from the aggregated energy signal of House 1 from UK-DALE dataset during the first 5 months of 2014 and with sampling rate 6s. As it is known from Word2Vec the created embeddings generalize much better when trained with as much data as possible. However, due to limited energy data and to demonstrate the transfer learning property, the training set of the embeddings excludes the rest of the UK-DALE data and all the REDD dataset.

The machine learning models that are selected for the evaluation are limited to lightweight models supporting multi-label classification. These models include decision trees, extra trees, random forests and feed-forward neural networks. The parameters of the models were selected using grid search. Among these classifiers the best ones were mostly variations of the neural network. In this series of experiments we avoid using any advanced or complex algorithms like ensembles to keep the time complexity as low as possible.

As far as the evaluation is concerned the metrics that are used in multi label classification tasks are micro and macro f1 scores. Macro-averaging firstly measures the performance of each class and then takes the mean, whereas micro-averaging calculates the overall performance [33]. The main difference is that macro-f1 emphasizes low performance of infrequent classes. Both measures are taken into consideration, however macro-average is used as the main metric to compare different models.

The methodology of model selection comprises pairing of each of the time series representations with all the candidate classifiers, including the various parameters of both representation algorithms and classifiers. Overall a few thousand different combinations are taken into consideration to avoid underrating any of the eight algorithms under evaluation. The model selection experiments are based on House 1 of UK-DALE dataset during the first half of 2014 and use a 5 cross-validation. The best machine learning models for each representation are presented in tables 1 and 2.

Power disaggregation depends a lot on the sampling rate or the length of the time frame that is used for inference. Despite the sampling rate is kept stable at 6s, different models and parameters are selected for different time frames. The length of the time frames varies from 10 minutes to one day. Regarding Signal2Vec, the initial embeddings, that have been created at the sampling rate of 6 seconds, are not retrained. A retrain would be desirable

Table 1: Model selection per time-frame for Signal2Vec, BOSS, DFT and PAA

Period	Signal2Vec	BOSS	DFT	PAA
10mins	NN(1000)	NN(2000,100,100)	ExtraTrees(500)	ExtraTrees(500)
30mins	NN(1000,100)	NN(2000,100,100)	ExtraTrees(500)	ExtraTree
1h	NN(1000,100)	NN(1000,100)	ExtraTrees(1000)	NN(2000,100,100)
2h	NN(1000,100)	NN(2000,100)	ExtraTrees(1000)	NN(100,100,100)
4h	NN(1000,100)	NN(1000,100)	ExtraTrees(200)	ExtraTrees(200)
8h	NN(1000)	NN(2000,100)	ExtraTrees(2000)	ExtraTrees(200)
24h	NN(1000,100)	NN(2000,100)	ExtraTrees(2000)	NN(1000,100))

Table 2: Model selection per time-frame for SFA, 1d-SAX, SAX and WEASEL

Period	SFA	1d-SAX	SAX	WEASEL
10mins	NN(2000,100,100)	RandomForest(100)	NN(2000)	NN(1000)
30mins	NN(2000,100,100)	RandomForest(100)	NN(2000)	NN(1000)
1h	NN(1000,2000,100)	ExtraTrees(100)	NN(100,)	NN(1000)
2h	NN(2000,100,100)	ExtraTrees(100)	NN(100,)	NN(100)
4h	ExtraTrees(500)	NN(1000,)	NN(1000,)	NN(100)
8h	NN(100,50,100,50)	NN(100,100)	NN(2000)	NN(2000,100)
24h	ExtraTrees(1000)	NN(1000,)	NN(100,100)	NN(1000,2000)

though, as it is expected to capture different frequency features and increase the performance. The only parameter that is configurable for Signal2Vec is the number of average vectors per time-frame. It has been found that longer time-frames require more vectors. This is not a surprise as very long signals leads to summing many vectors which in turn results in obscured information. The rest of the algorithms are parameterized using grid search. For more details of the parameters please refer to the related github repository. The appliances that are used during model selection are: oven, dish washer, fridge, washer dryer, boiler, vacuum cleaner, microwave, kettle, toaster, television, hair dryer and light.

5.2 Evaluation on UK-DALE

After model selection, the best machine learning models paired with the respective parameterized time series approximators are trained and tested. Training occurs in house 1 of UK-DALE dataset during March of 2013 and May of 2014. Testing takes place at the same house for the 12 appliances during the period June of 2014 and December of 2014. This is a very long period of testing, in contrast to the majority of the experiments in the literature, where NILM systems are tested for a period of a week or a couple of months. The results are presented in Figure 5. The two most efficient time series representations are BOSS and Signal2Vec, with the second one demonstrating overall the best performance. The following two models are DFT and PAA, showing quite robust results, especially when the time window is long.

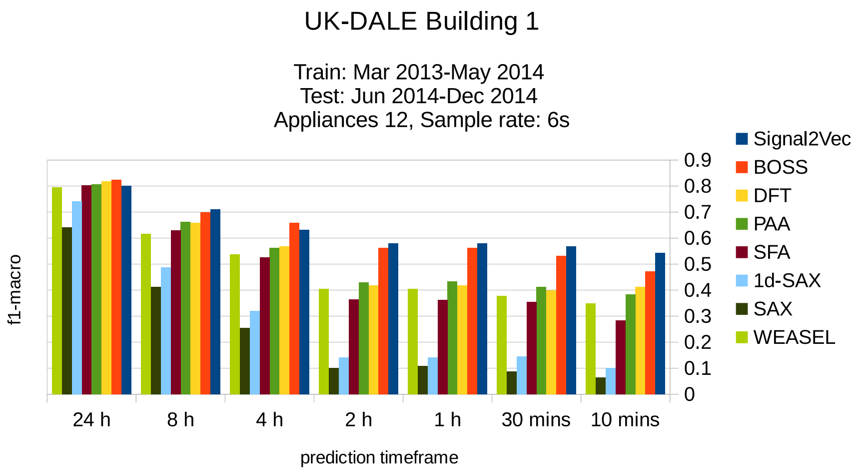


Fig. 5: Appliances: oven, microwave, dish washer, fridge, kettle, washer dryer, toaster, boiler, television, hair dryer, vacuum cleaner, light.

It is worth mentioning that Signal2Vec is the only representation which supports transfer learning. It was trained with a time-frame equal to 6 seconds and the same vectors are used for windows with duration from 10 mins to one day. All other representations are adapted and parameterized for each different time frame. The only parameter of Signal2Vec is the number of average vectors per time frame which is 1 for short windows and 5 for the ones longer than 2 hours. Another notable observation is that Signal2Vec not only outperforms BOSS, but also uses a smaller neural network with almost half neurons.

Another important parameter in energy disaggregation is the number of devices that are recognized. This set of experiments include 4 to 12 different appliances. Figure 6 presents some representative results with time period one hour. The 12 appliances are the ones that are previously mentioned. The set of 9 appliances contains: microwave, dish washer, fridge, kettle, washer dryer, toaster, television, hair dryer and vacuum cleaner. The set of 6 is: oven, dish washer, fridge, microwave, kettle, and toaster. Finally the set of 4 devices includes the ones that are mostly used in NILM research papers: dish washer, fridge, microwave and kettle. Signal2vec again achieves the best f1 macro score, with BOSS following. PAA, DFT and WEASEL are very close, SFA follows showing average performance, while SAX and 1d-SAX give the worse results.

Additionally, table 3 presents both f1 macro and micro, testing the disaggregation capabilities of two different sets of appliances of building 1 in UK-DALE dataset. One set includes high energy consumption devices such as: oven, dish washer, fridge, washer dryer, boiler and vacuum cleaner. The other set includes low energy consumption devices such as: microwave, kettle, toaster, television, hair dryer and light. As it is seen from the table, f1 macro

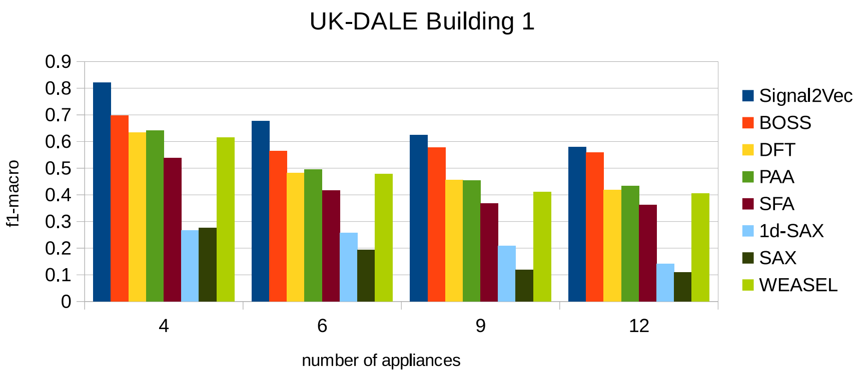


Fig. 6: Macro f1 score per number of appliances, for experiments with 1 hour time frame.

Table 3: Results disaggregating high and low cost appliances.

TS Repr.	High Cost f1-macro	High Cost f1-micro	Low Cost f1-macro	Low Cost f1-micro
Signal2Vec	0.55 ± 0.004	0.77 ± 0.008	**0.61 ± 0.008**	**0.73 ± 0.012**
BOSS	**0.56 ± 0.003**	**0.79 ± 0.001**	0.57 ± 0.002	0.67 ± 0.004
PAA	0.42 ± 0.015	0.77 ± 0.010	0.44 ± 0.010	0.55 ± 0.009
DFT	0.41 ± 0.003	0.78 ± 0.001	0.44 ± 0.001	0.60 ± 0.001
WEASEL	0.42 ± 0.000	0.76 ± 0.003	0.41 ± 0.001	0.52 ± 0.004
SFA	0.35 ± 0.007	0.73 ± 0.005	0.38 ± 0.011	0.52 ± 0.014
1d-SAX	0.19 ± 0.000	0.74 ± 0.003	0.07 ± 0.004	0.22 ± 0.010
SAX	0.20 ± 0.031	0.69 ± 0.078	0.03 ± 0.053	0.12 ± 0.214

and f1 micro most of the times are in good agreement. For the most costly appliances this time BOSS gives slightly better results than Signal2Vec, whereas for the low cost ones Signal2Vec outperforms BOSS.

5.3 Comparison with other multi-label NILM systems

The proposed lightweight multi-label NILM solutions are also evaluated on REDD dataset. Table 4 compares the performance of BOSS and Signal2Vec combined with a shallow multilayer perceptron against more computationally expensive models like RAkEL with an SVM base classifier and MLkNN [45]. The appliances that are disggregated in House 1 are: oven, refrigerator, light, microwave, bath GFI, outlet and washer. For House 3 the appliances are: electronics, furnace, washer dryer, microwave, bath GFI and kitchen outlet. The NILM solution that is suggested by Tabatabaei et al. uses features in the time or the frequency domain. In the time domain the method that is used is time delay embeddings and House 1 is best embedded with 8 dimensions

Table 4: Evaluation of multi-label NILM on REDD

House	RAkEL Time	MLkNN Time	RAkEL Wavelet	MLkNN Wavelet	NN Signal2Vec	NN BOSS
1	0.393	0.619	0.430	0.524	0.524	0.446
3	0.492	0.471	0.455	0.427	0.462	0.214

and a delay of 32 seconds, whereas house 3 with 18 dimensions and a delay of 95 seconds. In the frequency domain the signal is transformed using the Haar wavelet. According to the authors, the selected wavelet coefficients retain no less than 95% of the signal energy. Concerning the setup of BOSS and Signal2vec, the models are evaluated using 5 fold cross-validation. Inference period is 1 minute and 10 minutes, but without any significant difference in performance. The sampling rate is kept the same as for RAkEL and MLkNN at 1 second.

As shown in table 4 both BOSS and Signal2Vec show similar performance with the other models for House 1. Regarding House 3, BOSS underperforms, whereas Signal2Vec again shows very promising results. In overall, Signal2Vec surpasses wavelet based models in both houses. It is very important to emphasize that the vector space that is used by Signal2Vec, for the experiments on REDD dataset, is the same that is trained on UK-DALE dataset. Thus, transfer learning is not restricted in one dataset as the learnt embeddings generalize well for energy data from different countries.

6 Conclusion

This paper revisits the problem of power disaggregation through the lens of multi-label classification task. It is the first time that eight popular time series representations have been used in NILM and the experiments show that reducing the dimensions reduces the cost of a NILM solution and makes it possible to be contained in an embedded device. Among the time series representations, Signal2Vec demonstrates the most promising results, while it enables knowledge transferability. Finally, the proposed solutions are also compared with two other popular multi-label NILM solutions in the literature, showing equivalent performance and overpassing wavelet based models.

The exponential increase of connected devices and the data explosion problem make it mandatory to look for low cost solutions that can run on low end devices. Future research will focus on improving the framework of Signal2Vec and increase the performance of multi-label NILM systems.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. Agrawal, R., Faloutsos, C., Swami, A.: Efficient similarity search in sequence databases. In: International conference on foundations of data organization and algorithms, pp. 69–84. Springer (1993)
2. Armel, K.C., Gupta, A., Shrimali, G., Albert, A.: Is disaggregation the holy grail of energy efficiency? the case of electricity. Energy Policy **52**, 213–234 (2013)
3. Baranski, M., Voss, J.: Nonintrusive appliance load monitoring based on an optical sensor. In: 2003 IEEE Bologna Power Tech Conference Proceedings,, vol. 4, pp. 8–pp. IEEE (2003)
4. Basu, K., Debusschere, V., Bacha, S.: Load identification from power recordings at meter panel in residential households. In: 2012 XXth International Conference on Electrical Machines, pp. 2098–2104. IEEE (2012)
5. Basu, K., Debusschere, V., Bacha, S.: Residential appliance identification and future usage prediction from smart meter. In: IECON 2013-39th Annual Conference of the IEEE Industrial Electronics Society, pp. 4994–4999. IEEE (2013)
6. Basu, K., Debusschere, V., Bacha, S., Maulik, U., Bondyopadhyay, S.: Nonintrusive load monitoring: A temporal multilabel classification approach. IEEE Transactions on industrial informatics **11**(1), 262–270 (2014)
7. Basu, K., Debusschere, V., Douzal-Chouakria, A., Bacha, S.: Time series distance-based methods for non-intrusive load monitoring in residential buildings. Energy and Buildings **96**, 109–117 (2015)
8. Batra, N., Kelly, J., Parson, O., Dutta, H., Knottenbelt, W., Rogers, A., Singh, A., Srivastava, M.: Nilmtk: an open source toolkit for non-intrusive load monitoring. In: Proceedings of the 5th international conference on Future energy systems, pp. 265–276. ACM (2014)
9. Batra, N., Kukunuri, R., Pandey, A., Malakar, R., Kumar, R., Krystalakos, O., Zhong, M., Meira, P., Parson, O.: Towards reproducible state-of-the-art energy disaggregation. In: Proceedings of the 6th ACM International Conference on Embedded Systems for Energy-Efficient Built Environments (BuildSys' 19). ACM, New York, NY, USA (2019)
10. Chang, H.H., Chien, P.C., Lin, L.S., Chen, N.: Feature extraction of non-intrusive load-monitoring system using genetic algorithm in smart meters. In: 2011 IEEE 8th International Conference on e-Business Engineering, pp. 299–304. IEEE (2011)
11. Egarter, D., Pöchacker, M., Elmenreich, W.: Complexity of power draws for load disaggregation. arXiv preprint arXiv:1501.02954 (2015)
12. Froehlich, J., Larson, E., Gupta, S., Cohn, G., Reynolds, M., Patel, S.: Disaggregated end-use energy sensing for the smart grid. IEEE Pervasive Computing **10**(1), 28–39 (2010)
13. Gutmann, M.U., Hyvärinen, A.: Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. Journal of Machine Learning Research **13**(Feb), 307–361 (2012)
14. Hart, G.W.: Nonintrusive appliance load monitoring. Proceedings of the IEEE **80**(12), 1870–1891 (1992). DOI 10.1109/5.192069
15. Kato, T., Cho, H.S., Lee, D., Toyomura, T., Yamazaki, T.: Appliance recognition from electric current signals for information-energy integrated network in home environments. In: International Conference on Smart Homes and Health Telematics, pp. 150–157. Springer (2009)
16. Kelly, J., Knottenbelt, W.: The uk-dale dataset, domestic appliance-level electricity demand and whole-house demand from five uk homes. Scientific data **2**, 150007 (2015)
17. Keogh, E., Chakrabarti, K., Pazzani, M., Mehrotra, S.: Dimensionality reduction for fast similarity search in large time series databases. Knowledge and information Systems **3**(3), 263–286 (2001)
18. Kolter, J.Z., Johnson, M.J.: Redd: A public data set for energy disaggregation research. In: Workshop on Data Mining Applications in Sustainability (SIGKDD), San Diego, CA, vol. 25, pp. 59–62 (2011)
19. Krystalakos, O., Nalmpantis, C., Vrakas, D.: Sliding window approach for online energy disaggregation using artificial neural networks. In: Proceedings of the 10th Hellenic Conference on Artificial Intelligence, p. 7. ACM (2018)

20. Lai, Y.X., Lai, C.F., Huang, Y.M., Chao, H.C.: Multi-appliance recognition system with hybrid svm/gmm classifier in ubiquitous smart home. Information Sciences **230**, 39–55 (2013)
21. Laughman, C., Lee, K., Cox, R., Shaw, S., Leeb, S., Norford, L., Armstrong, P.: Power signature analysis. IEEE power and energy magazine **1**(2), 56–63 (2003)
22. Li, D., Dick, S.: Whole-house non-intrusive appliance load monitoring via multi-label classification. In: 2016 International Joint Conference on Neural Networks (IJCNN), pp. 2749–2755. IEEE (2016)
23. Li, D., Dick, S.: Residential household non-intrusive load monitoring via graph-based multi-label semi-supervised learning. IEEE Transactions on Smart Grid (2018)
24. Li, D., Sawyer, K., Dick, S.: Disaggregating household loads via semi-supervised multi-label classification. In: 2015 Annual Conference of the North American Fuzzy Information Processing Society (NAFIPS) held jointly with 2015 5th World Conference on Soft Computing (WConSC), pp. 1–5. IEEE (2015)
25. Liang, J., Ng, S.K., Kendall, G., Cheng, J.W.: Load signature study—part i: Basic concept, structure, and methodology. IEEE transactions on power Delivery **25**(2), 551–560 (2009)
26. Lin, J., Keogh, E., Wei, L., Lonardi, S.: Experiencing sax: a novel symbolic representation of time series. Data Mining and knowledge discovery **15**(2), 107–144 (2007)
27. Lin, J., Khade, R., Li, Y.: Rotation-invariant similarity in time series using bag-of-patterns representation. Journal of Intelligent Information Systems **39**(2), 287–315 (2012)
28. Malinowski, S., Guyet, T., Quiniou, R., Tavenard, R.: 1d-sax: A novel symbolic representation for time series. In: International Symposium on Intelligent Data Analysis, pp. 273–284. Springer (2013)
29. Marchiori, A., Hakkarinen, D., Han, Q., Earle, L.: Circuit-level load monitoring for household energy management. IEEE Pervasive Computing **10**(1), 40–48 (2010)
30. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems, pp. 3111–3119 (2013)
31. Nalmpantis, C., Vrakas, D.: Machine learning approaches for non-intrusive load monitoring: from qualitative to quantitative comparation. Artificial Intelligence Review **52**(1), 217–243 (2019)
32. Nalmpantis, C., Vrakas, D.: Signal2vec: Time series embedding representation. In: International Conference on Engineering Applications of Neural Networks, pp. 80–90. Springer (2019)
33. Rak, R., Kurgan, L., Reformat, M.: Multi-label associative classification of medical documents from medline. In: Fourth International Conference on Machine Learning and Applications (ICMLA'05), pp. 8–pp. IEEE (2005)
34. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. Machine learning **85**(3), 333 (2011)
35. Restuccia, F., D'Oro, S., Melodia, T.: Securing the internet of things in the age of machine learning and software-defined networking. IEEE Internet of Things Journal **5**(6), 4829–4842 (2018)
36. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics **20**, 53–65 (1987)
37. Ruzzelli, A.G., Nicolas, C., Schoofs, A., O'Hare, G.M.: Real-time recognition and profiling of appliances through a single electricity sensor. In: 2010 7th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON), pp. 1–9. IEEE (2010)
38. Schäfer, P.: The boss is concerned with time series classification in the presence of noise. Data Mining and Knowledge Discovery **29**(6), 1505–1530 (2015)
39. Schäfer, P., Högqvist, M.: Sfa: a symbolic fourier approximation and index for similarity search in high dimensional datasets. In: Proceedings of the 15th International Conference on Extending Database Technology, pp. 516–527. ACM (2012)
40. Schäfer, P., Leser, U.: Fast and accurate time series classification with weasel. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 637–646. ACM (2017)

41. Singh, S., Majumdar, A.: Non-intrusive load monitoring via multi-label sparse representation based classification. IEEE Transactions on Smart Grid (2019)
42. Singhal, V., Maggu, J., Majumdar, A.: Simultaneous detection of multiple appliances from smart-meter measurements via multi-label consistent deep dictionary learning and deep transform learning. IEEE Transactions on Smart Grid **10**(3), 2969–2978 (2018)
43. Srinivasan, D., Ng, W., Liew, A.: Neural-network-based signature recognition for harmonic source identification. IEEE Transactions on Power Delivery **21**(1), 398–405 (2005)
44. Suzuki, K., Inagaki, S., Suzuki, T., Nakamura, H., Ito, K.: Nonintrusive appliance load monitoring based on integer programming. In: 2008 SICE Annual Conference, pp. 2742–2747. IEEE (2008)
45. Tabatabaei, S.M., Dick, S., Xu, W.: Toward non-intrusive load monitoring via multi-label classification. IEEE Transactions on Smart Grid **8**(1), 26–40 (2016)
46. Takens, F.: Detecting strange attractors in turbulence. In: Dynamical systems and turbulence, Warwick 1980, pp. 366–381. Springer (1981)
47. Tsoumakas, G., Katakis, I.: Multi-label classification: An overview. International Journal of Data Warehousing and Mining (IJDWM) **3**(3), 1–13 (2007)
48. Verma, S., Singh, S., Majumdar, A.: Multi label restricted boltzmann machine for non-intrusive load monitoring. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8345–8349. IEEE (2019)
49. Wu, X., Kumar, V.: The top ten algorithms in data mining. CRC press (2009)
50. Zhang, C., Zhong, M., Wang, Z., Goddard, N., Sutton, C.: Sequence-to-point learning with neural networks for non-intrusive load monitoring. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
51. Zhang, M.L., Zhou, Z.H.: Ml-knn: A lazy learning approach to multi-label learning. Pattern recognition **40**(7), 2038–2048 (2007)
52. Zhong, M., Goddard, N., Sutton, C.: Latent bayesian melding for integrating individual and population models. In: Advances in neural information processing systems, pp. 3618–3626 (2015)
53. Zia, T., Bruckner, D., Zaidi, A.: A hidden markov model based procedure for identifying household electric loads. In: IECON 2011-37th Annual Conference of the IEEE Industrial Electronics Society, pp. 3218–3223. IEEE (2011)