# Data analytics for home appliances identification

Ayush Garg, Gabriel Vizcaino and Pradeep Somi Ganeshbabu

Energy Science Technology and Policy

Carnegie Mellon University, Pittsburgh, PA 15213

**ABSTRACT**

In this paper, we will be identifying home appliances using the voltage and current data obtained from different smart meters. A public dataset of high resolution for load identification research called as PLAID was used. The voltage vs current profile is used as features to identify the appliances. The PLAID dataset is loaded and the current vs voltage profile for the 1074 appliances (11 unique types) is plotted for one cycle in steady state for each appliance. The images are then saved and loaded as a matrix. After using different classifiers, the Random Forest Classifier gave the best accuracy and time.

**CCS Concepts**

•

**Keywords**

Keywords are your own designated keywords separated by semicolons (“;”).

# INTRODUCTION

The energy consumption in the residential and commercial account for 41% of the total energy consumed in the United States [1]. This growth can be attributed to a list of appliances like ceiling fans, fume hoods, vending machines, televisions, computers etc. These loads are termed as Miscellaneous Electrical Loads (MELs). The primary energy consumption of the MELs is expected to grow from 6.1 quads to 6.9 quads in residential buildings and from 6.5 quads to 8.3 quads for commercial buildings [2]. The 2015 Annual Energy Outlook has projected a growth of 13% in residential MELs primary energy consumption from 2016 to 2030 and a 27% increase in commercial buildings during the same period [2]. Growth of MEL has resulted in offset of some of the efficiency gains and improvement in standards of major end use appliances like space conditioning, lighting, and water heating.

For this analysis, we have used the Plug-Load Appliance Identification Dataset(PLAID). PLAID is a public dataset. The current and voltage measurements are sampled at 30 kHz for 11 different appliance types. More than 200 different makes and models of the 11 different types are used, which are present in the 55 households in Pittsburgh. A similar work on the PLAID data is done by [3], achieving 86.03% average accuracy across different classifiers, when training and testing with different appliances.

# EXPERIMENT STEPS

The Experiment

Before doing, the analysis we considered some assumptions. Since the data consisted of both steady state and transient state data, we took the last 30,000 data of each appliance. After plotting the data, we found out that 505 data points make one cycle. We then found the total number of cycles, and took the mean of each point in each cycle. After finding the different type of appliances we have in data, we plotted the current vs voltage for five random appliances of each type. Based on these values we plotted the current and voltage profile of each appliance for the last ten steady state periods. This is show in figure 1.

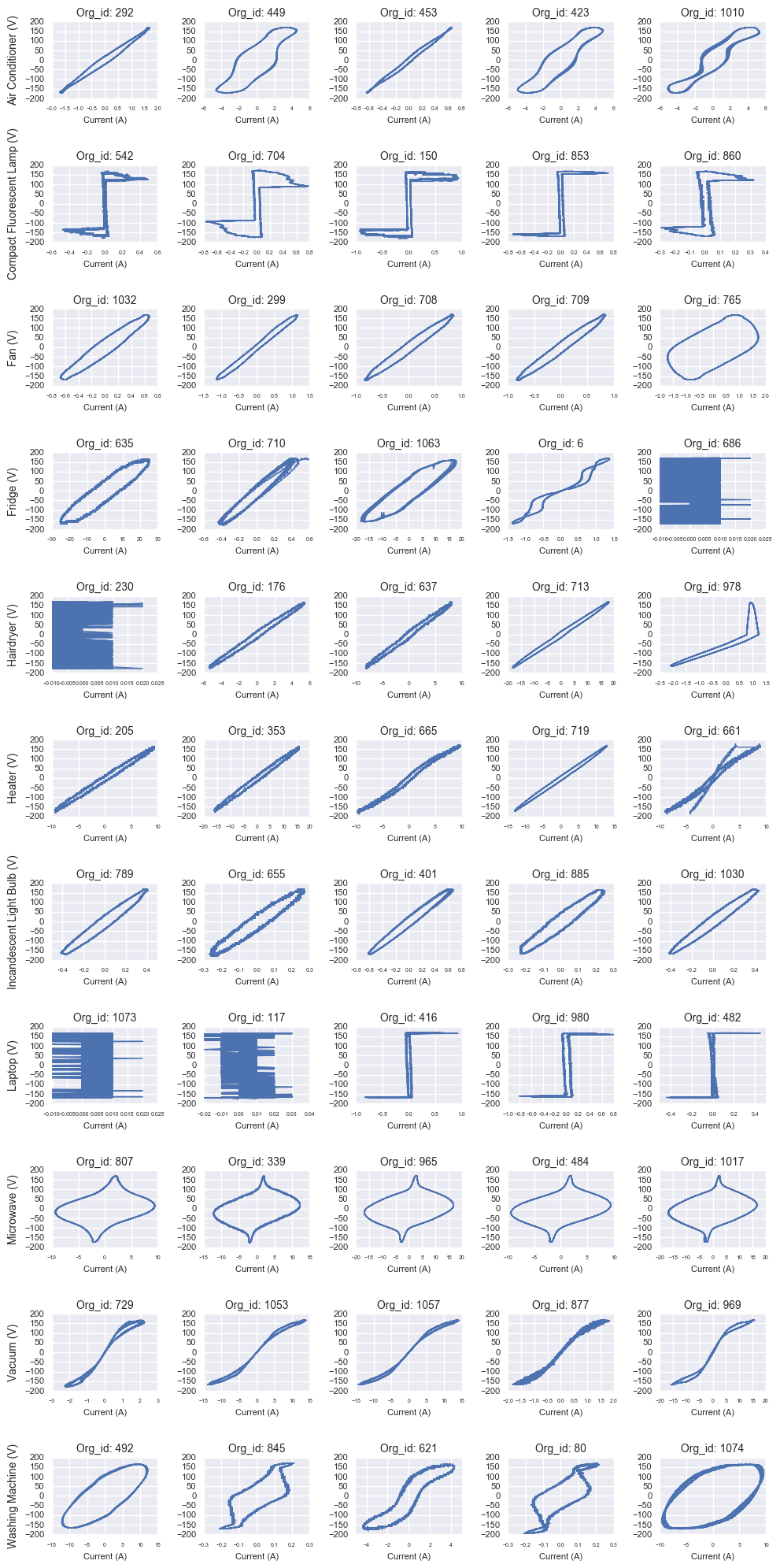
From the V-I plots above we can conclude that, especially in the steady state, the combination of linear and non-linear elements within each appliance type produces a similar pattern of voltage vs. current across appliances of the same type. Though not perfectly consistent, we can harness this characteristic in order to build features that help us classify an appliance given its voltage and currents signals.

We explored different transformations to extract features from voltage and current signals like directly using the voltage and current values, calculating the Fourier transform of the current to identify harmonics, descriptive statistics (e.g. standard deviations and variation coefficients over a cycle) and printing images of V-I plots in order to extract the pixels’ characteristics. While all of them provide useful information to identify appliances, the latter (i.e. images) is the transformation that yields the highest predicting accuracy. Therefore, we stick with this approach.

We also normalize the current and voltage data of the last cycle (assuming it is steady state) and print the current and voltage data as images for each individual appliance. After loading the images as arrays, with each image being 180 X 180 pixels, a matrix is creating which will become the matrix of features. The V-I pattern images saved as png files significantly use less memory that the raw data in csv files (~8 MB the whole folder).

To build a well-performing classifier to identify the appliance type based on its voltage and current signals as inputs, particularly the V-I profile at steady state, we start by evaluating different multi-class classifiers on the features matrix. The eight different classifiers used are one-vs-rest classifier, Extra Tree classifier, Decision Tree classifier, Gaussian NB, Bernouli NB, Gradient Boosting classifier, k-Neighbors Classifier, Random Forest classifier, and Random guess classifier. To prevent overfitting, the dataset is randomly divided into three sub-sets: training, validation, and test. The models are fitted using the training subset and then the accuracy is tested on the validation subset. After this evaluation the best models are fine-tuned and then tested using the testing subset. Since the objective is to accurately identify the type of an appliance based on its electrical signals, the following formula is used to measure accuracy [4]:

Accuracy (Score) =



**Figure 1 V-I curve for last ten steady state periods**

# RESULTS

In general, the evaluated classifiers remarkably improve over the default classifier - expect for the Naive Bayes classifier using Bernoulli distributions (as expected given the input data). The one-vs-the-rest model, using a support vector machine estimator, is the one showing the highest accuracy on the validation subset. However, this classier, along with the Gradient Boosting (which also presents a good performance), takes significantly more time to fit than the others. On the contrary, the K-nearest-neighbors and Random Forest classifiers also achieve high accuracy but much faster. For these reasons, we are going to fine tune the main parameters of the latter two classifiers, re-train them, and then test again their performance on the testing subset. Table 1 shows the results of the various classifiers used.

Table 1. Different classifier accuracy and time

|  |  |  |
| --- | --- | --- |
| **Classifier** | **Accuracy** | **Time(s)** |
| One-Vs-Rest | 0.9012 | 262.054 |
| Extra Tree | 0.7326 | 0.144 |
| Decision Tree | 0.8023 | 1.642 |
| Gaussian NB | 0.5058 | 1.566 |
| Bernouli NB | 0.157 | 0.551 |
| Gradient Boosting | 0.843 | 822.472 |
| k-Neighbor | 0.8779 | 8.883 |
| Random Forest | 0.8547 | 0.369 |

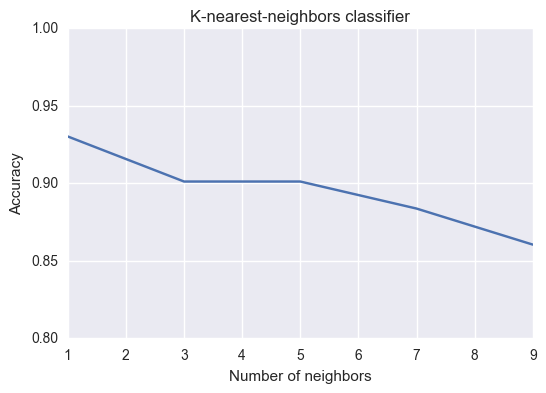
## Parameter tuning

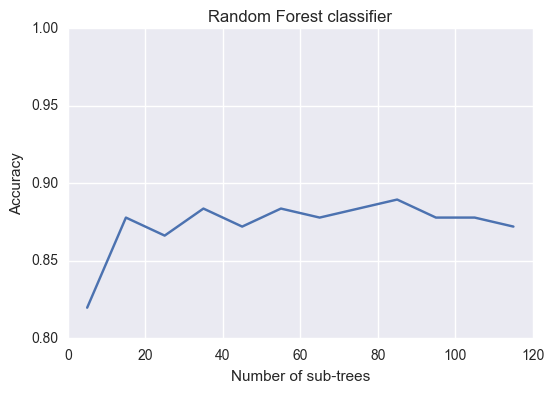
For the KNN classifier, the above graph suggests that the less number of neighbors to consider, the better the accuracy. Therefore, we are going to set this parameter to have only one neighbor in the KNN classifier. Having this new parameters, the training and validation sub-sets were retrained for both the classifiers, and test the fitted model on the testing set. Figure 2 shows us how the accuracy changes with number of neighbors.

Although the characteristic of the Random Forest classifier entails that the shape of the above graph changes every time it is run, the general behavior suggests that having more than 10 sub-trees notably improves the performance of the classifier. Progressively increasing the number of trees after this threshold slightly improves the performance further, up to a point, around 70-90, when the accuracy starts decreasing. Therefore, the number of parameters is limited to 80 sub-trees. Figure 3 Table 2 shows the results after fine tuning the parameters.

Table 2. Classifier results after fine tuning parameters

|  |  |  |
| --- | --- | --- |
| **Classifier** | **Accuracy** | **Time(s)** |
| k-Neighbor | 0.9116 | 9.546 |
| Random Forest | 0.884 | 1.468 |



**Figure.2-Change in accuracy with changing number of neighbors** 

**Figure.3-Change in accuracy with changing number of subtrees**

Both classifiers improved their performance after the tuning their parameters. KNN even outweighs the performance of the one-vs-the-rest classifier. Although the score of the Random Forest classifier slightly lags behind KNN, this fitting time of this one is 8x times faster than KNN. To further test the performance of both classifiers, a random 10-fold cross-validation process was used on both models using the whole dataset.

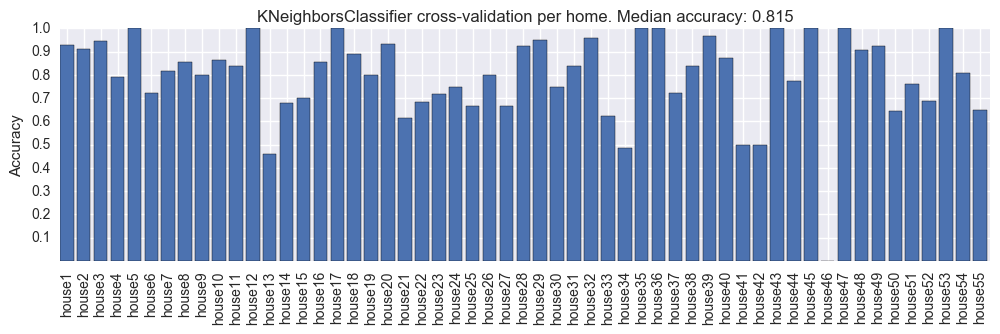
Table 3. Classifier results after fine tuning parameters

|  |  |  |
| --- | --- | --- |
| **Classifier** | **Average score** | **10-fold CV time(s)** |
| k-Neighbor | 0.934 | 76.215 |
| Random Forest | 0.92 | 20.065 |

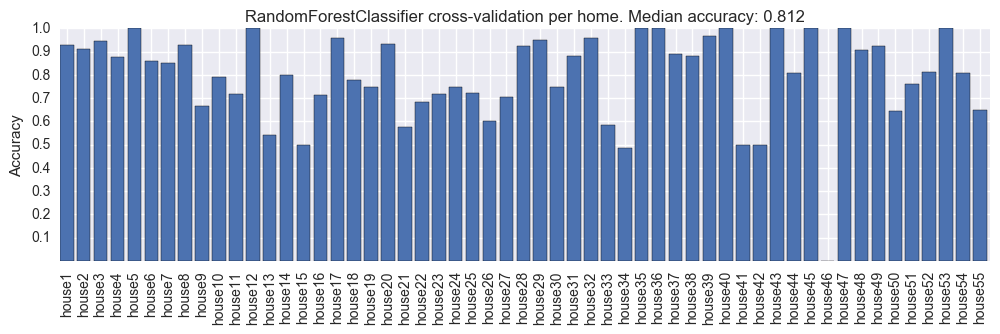
The results from the 10-fold cross-validation are very promising. Both models present more than 92% average accuracy and though KNN scores slightly higher, the Random Forest still shows significantly lesser fitting time.

## Identifying appliance type per house

One last step to test the performance of the KNN and Random Forest classifiers would be to predict or identify the type of appliances in particular house, based on the voltage and current signals, by training the model on the data from the rest of the houses. There are 55 homes surveyed and each appliance has a label indicating its corresponding house; hence, it is possible to split the data in this fashion. This is another kind of cross-validation.



**Figure 4- k-Neigbours classification per house**

****

**Figure 5 – Random Forest classifier per house**

Figure 4 and Figure 5 show results of the cross-validation per home show a median accuracy above 80% for both classifiers. Out of the 55 home appliance predictions, 9 scored 100% accuracy and around 20 had scores above 90%. Only 3 and 2 houses had a scored below 50% using KNN and RF respectively. In general, the presented outcome suggests that the chosen classifiers work fairly well, although they perform poorly for certain homes. In order to identify why is this the case, it is worth it to plot the predictions and actual type of a couple of those home appliances.

# CONCLUSIONS AND FUTURE WORK

This paper presents a data-driven approach to the problem of identifying home appliances type based on their corresponding electrical signals. Different multi-class classifiers are trained and tested on the PLAID dataset in order to identify the most accurate and less computationally expensive models. An image recognition approach of Voltage-Current profiles in steady state is used to model the inputs of the appliance classifiers. Based on the analyses undertaken we are able to identify some common patterns and draw conclusions about the two best performed classifiers identified in terms of time and accuracy, k-nearest-neighbors and Random Forest Decision Tree:

* After fine tuning their corresponding parameters on a training sub-set, the average accuracy of KNN and RF, applying 10-fold cross-validation, is greater than 91%.
* The One-vs-the-rest and Gradient Boosting Decision Trees classifiers also show high accuracy; however, the fitting time is in the order of minutes (almost 15 min. for Gradient Boosting), whereas KNN and RF take seconds to do the job.
* Though KNN scores slightly higher than RF, the latter takes significantly shorter fitting time (about 8x time less).
* While high accuracy in both classifiers is achieved using traditional cross-validation techniques, when applying cross-validation per individual home, the accuracy decreased to 80% on average.
* While debugging the classifiers we noticed that many of the input signals of current and voltage do not reach steady state in different appliances. Therefore, their corresponding V-I profile is not well defined which makes the prediction harder even for a human expert eye. We also noticed that in several homes, the list of associated appliances contain the same appliance sampled in different times. Therefore, in those cases the classifiers are meant to failed repeatedly in a single house.

The following task are proposed as future work in order to improve the performance of the trained appliance classifiers:

* Collect more data: The figure bellow shows the training and test accuracy evolution of the RF classifier with respect to the number of samples. While only slight increments are realized after 700-800 samples, it seems that there is still room for improvement in this sense.

# REFERENCES

1. Sameer Kwatra, Jennifer Amann, and Harvey Sachs, “Miscellaneous energy loads in Buildinds”, American Council for an Energy- efficient economy (June 2013)
2. “DOE seeking information on miscellaneous electrical loads research and development topics” <http://energy.gov/eere/buildings/articles/doe-seeking-information-miscellaneous-electrical-loads-research-and> [Online] Available (Accessed 12/8/2016)
3. J. Gao, S. Giri, E. C. Kara, and M. Berg´es, “Plaid: A public dataset of high-resolution electrical appliance measurements for load identiﬁcation research: Demo abstract,” in Proceedings of the 1stACM Conference on Embedded Systems for Energy-Efﬁcient Buildings, BuildSys ’14, (New York, NY, USA), pp. 198–199, ACM, 2014.
4. Scikit learn “Machine learning in Python” <http://scikit-learn.org/stable/> [Online] Available (Accessed 12/09/2016)