**Machine Learning for Identification and Generalization of Appliances in Smart Homes**

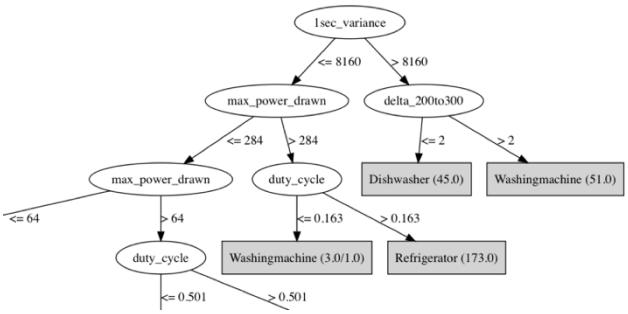
Tucker Williams, Class of 2018

This summer I worked with Professor Barker on using machine learning to identify appliances in a smart home. Using mainly data from *Tracebase* and the *Weka* toolkit, we specifically applied my efforts towards identifying unseen devices.

In this context, a smart home is a collection of data files containing one-second power readings (e.g., *12:00:01-12:00:02;45* means a 45 watt output for that second) for all classes of appliances (e.g., Refrigerator, Microwave, …) and for all specific devices in those appliance classes (e.g., Samsung_Refrigerator_A123.)

By previously unseen devices, we mean that we would like to train on Samsung_Refrigerator_A123 and *not* train on but still identify GE_Refrigerator_B234. Can we generalize models such that we can identify devices we've never seen before – whether Samsung, GE, LG, etc.? This is a focus of our work.

For a baseline in our project, we confined our experimentation to three arbitrarily chosen but fundamental algorithms: Naïve Bayes, Support Vector Machines, and Decision Tree classifiers. Towards the end of the project we primarily focused on Decision Trees for their intuitive, human-understandable tree structure (see below.)




At left is an example decision tree output. Based on features in the data (which we calculate using the 1-second power readings), the algorithm first trains on seen data, meaning data that is labeled as "Washing_Machine," and then attempts to test on unseen data, meaning data that is not labeled but still has the features. We hope to identify correctly given those features.

I spent a good portion of my time during the eight weeks building a project infrastructure that coincided with the data. This infrastructure required processing the data, identifying characteristics in the data, building specific files that work with Weka, and writing programs that would automate and report on experiment results.

Around halfway through the eight weeks, we shifted our experimental focus (concurrent with building out the project infrastructure) more and more to identifying unseen devices. It is easy to be accurate with many of these machine learning algorithms if you have *all* the data, but that is never the case is the real world and thus neither as interesting nor applicable.

Once we shifted focus, I wrote a program to run automatic experiments on this schema of seen/unseen devices. Accuracy remained quite low for unseen devices, though it improved with the number of devices seen (i.e., if there are five devices overall, and we test on all five always, training on one device, then two, then three, so on up to five.) The question then becomes: to improve accuracy, do we need more or better features or is it simply a question of data-scale? As said earlier, the results were very promising with *all* the data; so, presumably, this accuracy would improve in the real world as the algorithm received more data – which was experimentally shown by accuracy improving with the proportion of devices trained.

In the end, we verified the intuitive idea that accuracy improves with exposure to more data. It remains to be seen, however, how it would fair in the real world of millions of different devices. This is currently a problem in smart home research – the general lack of data. It is currently accepted dogma that proper AI and machine learning requires reams of data on the scale of terabytes and petabytes. Right now, this is not something available to smart home research.

**Faculty Mentor**: Sean Barker
**Funded by the**: Maine Space Grant Consortium
**Data sources:** https://www.tracebase.org/traces