Interim Report

Sam Stern (s1134468)

February 9, 2015

Aim of the project and goals

Amidst international pressure on countries to reduce their carbon footprint [3] and the British public becoming increasingly frustrated by rising energy bills will little explanation to the cause in this rising price [6], the UK Government is currently executing on a plan to distribute smart meters to households accross the country by 2020. Smart meters, which measure a household's gas and electricity consumption in real-time, are expected to both help a household reduce its energy usage by displaying how much energy is being used , as well as increase transparency in the household's bills by eliminating the need for monthly meter readings and estimations by the energy companies. Instead, the energy providers are given a much more accurate description of the household's consumption and as a result, will be able give a more accurate bill.

While there has generally been strong support for the smart meter program, there has also been resistance to the campaign with fears that the energy companies will use the information as an opportunity to raise their customers bills and increase their own profit [7]. Perhaps more interestingly though, and therefore the focus of this project, are the concerns which have been have been raised regarding the risk associated with measuring and storing energy consumption data [4] [5]. Particularly, to what extent can other information about a household be inferred from energy consumption readings?

The aim of this project is to explore whether (and to what extent) it is possible to construct features that can be used to predict detailed personal information of a household from their energy consumption readings, by taking on the role of a malicious individual (or group) who wishes to exploit this information to determine household properties, which will be referred to as classes, that might be of interest to someone wishing to either target advertise or burgle a household. Using household electricity consumption information collected by the Household Electricity Survey (HES), a DEFRA sponsored national survey of energy use collected over a period from 2010 to 2011, classification models are created to predict two household properties: (1) The presence (or absence) of children in a household and (2) the IPSOS social grade of the chief income earner of the household. These properties are chosen because, of the information gathered by the HES survey, they are of logical interest to someone who might wish to intrude on a household.

This project has 3 main components:

1. Clean the data and create a database that stores the house sets and relevant household and energy-use information
2. Find useful features from the data that can be used as inputs to a logistic regression model
3. Predict household properties using supervised learning methods

Results and Accomplishments so far

Cleaning The Data

Rather than measuring the total energy used by a household in a given time interval, as is required for this task, the HES data contains readings of individual appliances and sockets of the households. To get an estimate of the mains reading, the measurements for individual appliances need to be aggregated together. Cambridge Architectural Research Limited edited much of the data and created a list that maps the appliances that need to be aggregated for each household, in order to get an estimate of the mains reading. After the mains reading is created, several additional steps are taken to pre-process the data before feature construction and classification can take place.

While some households had readings taken with in two minute intervals, others only have a granularity of 10 minutes. To make the data uniform, every 5 readings of the 'two-minute households' are summed, resulting in all households having 10-minute granularity.

As the smart meters were installed (and uninstalled) in different households at different times of day, the data is topped and tailed so that each instance starts and ends at midnight, ensuring that each instance has contains an integer number of days.

The next task is to ensure that all instances are of equal length. Of the 250 households that participated in the study, 26 had measurements taken for 1 year, while the remaining 224 had readings taken anywhere from 24 to 31 days. The initial approach was take each year-long household and split it into 12 month-long households, resulting (in effect) in 536 household instances. From there, the shortest instance was found (24 day) and the tail of each household was chopped off to 24 days (or 3456 dimensional data).

While this does result in a dataset that could be used as an input for a classification model, there are some obvious concerns that can be raised. The first one being: Is it necessary to eliminate so much data? While some households had their electricity read for as few as 24 days, most were measured for closer to 28 days, resulting in an average loss of 4 days or 576 meter readings (if we exclude the households that were measured for a year). Furthermore, visualizing the data indicated that in addition to there being obvious periodicity over each day, many households also exhibit periodicity over a week too. Chopping the data as described above does not capture this pereodicity and fails to account account for any phase shift that would effect frequency domain related features [8].

The solution is to create four-week (28 day) instances. Where the data is reused in households that are less than 28 days long. This is done in in the following steps:

- For the month-long households
    1. Find the mode (most commonly appearing) day of the week that the measurements begin on (this turns out to be a Sunday).
    2. Chop the tops off the household instances that don't start on a Sunday (but don't discard as this can often be recycled)
    3. Append days to the end until there is 4 weeks forth of data, ensuring that a weekly cycle is maintained (i.e a data from a Tuesday can only follow that of a Monday, which can only follow a Sunday etc).
- For the year-long households
    1. Using the mode start day found earlier, group the data into 4 week periods.
    2. Treat each of these groups as individual households.

This method results in 520 instances (once outliers have been removed). The advantage to this method is that it retains more data than the first approach and ensure that the weekly periodicity is still captured.

The cleaning of the data and pre-processing was done by creating a MySQL database and using python with embedded to write scripts that read from and write to the database.

Feature Extraction

In order to both reduce dimensionality, as well as capture the distinction between different classes, the data from the electricity meter readings is manipulated in order to extreact useful features.

According to [2], possible features that are interesting for classification of households based on energy consumption are: consumption figures, ratios, temporal properties, and statistical properties. Consumption figures are the average, maximum and minimum energy consumption over some time period. Ratios are features that calculate the ratio between consumption figures and can capture relevant patterns that occure through different time intervals. Temporal features capture the first (or last) time some event takes place which or at what time the daily maximum occures. Finally, statistical properties such as variance, give insight into the consumption curve (for example how a households energy consumption correlates with itself. In addition to these, I also look at frequency domain related features, particularly the discrete fourier transform of the data.

Features Constructed thus far include:

- Total energy used over the 4 week period
  - Sum of all the smart meter readings for the household
- Average total energy used for each day of the week
  - Every 144 readings are summed together (6 readings/hour*24 hours/day=144 readings per day).The data is then grouped by day of the week and the average is taken
- Part-of-day features
  - Most schools in the UK start between 8:30 and 9:00 and end between 15:00 and 16:00. With this in mind, as well as with the assumption that children will go to bed before 22:00, days are split into four groups: 6:00-9:00,9:00-15:00,15:00-22:00 and 22:00-6:00. The features created consist of the sums of the grouped data.

- Average part-of-day for each day of the week
  - Takes the average of the part-of-day features described above (i.e the average Monday 6:00-9:00, or Thursday 15:00-22:00), as each of these will occur four times over the 28 day period and there is no reason to assume that the first Monday morning is different from the second Monday morning)
- Discrete Fourier Transform (DFT) of meter readings
  - Converts the data to a list of coefficients of a sinusoids, ordered by their frequencies [cite wiki]. The F-test ($F = \frac{variance\ between\ classes}{variance\ within\ classes}$) is then used to select the features which best discriminate between the classes.
- DFT ex-weekend
  - Remove weekend data before taking the DFT as energy use patterns may be different on the weekends and could therefore affect the resulting coefficients of the sinusoids.

Timeline

There are still four steps to complete before, each of which I estimate will take a week to complete:

- Compare features used to classify children vs. no-children households
- Adapt code for multi-class case as well as explore other features that may be relevant for socio-economic classification.
- Construct model for socio-economic classification
- Write final report

Report Outline

1. Introduction
   (a) Introduction
   (b) Smart Meters
   (c) Aim of the Project

    (d) Related Work

2. Data

    (a) Overview of HES Dataset
    (b) Extracting the Data and Pre-Prossessing
    (c) Issues Encountered With the Data
    (d) Comparison to Previous Work

3. Feature Exploration

    (a) Featured for Classifying Households With Children
    (b) Featured for Classifying Socie-Economic Status of Households
    (c) Feature Selection
    (d) Class Cardinality

4. Models

    (a) Overview and Rationale Behind Choosing Models
    (b) Logistic Regression
    (c) Multinomial Logistic Regression

5. Results

    (a) Presence/Absence of Children
    (b) IPSOS Socio-Economic Class

6. Conclusion and Further Work

    (a) Conclusion
    (b) Further Work

References

[1] Christian Beckel, Leyna Sadamori, and Silvia Santini. Automatic socio-economic classification of households using electricity consumption data. In *Proceedings of the Fourth International Conference on Future Energy Systems*, e-Energy ‘13, pages 75–86, New York, NY, USA, 2013. ACM.

[2] Christian Beckel, Leyna Sadamori, and Silvia Santini. Towards automatic classifi- cation of private households using electricity consumption data. In *Proceedings of the Fourth ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*, BuildSys '12, pages 169–176, New York, NY, USA, 2012. ACM.

[3] European commission. directive 2006/32/ec of the european parliament and of the council of 5 april 2006 on energy end-use efficiency and energy services and repealing council directive 93/76/eec, 2006.

[4] Elias Leake Quinn. Privacy and the new energy infrastructure (february 15, 2009). `http://ssrn.com/abstract=1370731` or `http://dx.doi.org/10.2139/ ssrn.1370731`.

[5] M.A. Lisovich, D.K. Mulligan, and S.B. Wicker. Inferring personal information from demand-response systems. *Security Privacy, IEEE*, 8(1):11–20, Jan 2010.

[6] Office for National Statiscics. 2014. Full Report: Household Energy Spending in the UK, 2002-2012. [ONLINE] Available at: `http://www.ons.gov.uk/ons/dcp171776_354637.pdf`. [Accessed 26 January 15].

[7] STOPSMARTMETERS. 2012. Stop Smart Meters. [ONLINE] Available at: `http://stopsmartmeters.org.uk/`. [Accessed 26 January 15].

[8] Smith Steven W. (1997) *The Scientist and Engineer's Guide to Digital Signal Processing* California Technical Pub