Occupancy Prediction using Building Energy Consumption Data

Carnegie Mellon University

Karthik Seeganahalli
Carnegie Mellon University
Civil and Environmental Engineering
5000 Forbes Avenue, Pittsburgh 15217
United States
kkrishn2@andrew.cmu.edu

Soumyajit Paul
Carnegie Mellon University
Civil and Environmental Engineering
5000 Forbes Avenue, Pittsburgh 15217
United States
soumyajp@andrew.cmu.edu

Joseph Chou
Carnegie Mellon University
Civil and Environmental Engineering
5000 Forbes Avenue, Pittsburgh 15217
United States
jchou1@andrew.cmu.edu

ABSTRACT

With economic upsurge, the energy demand and consumption pattern of built spaces is under a paradigm shift in today's world. The time variant energy pricing is one of the solution for effective energy consumption. The automated occupancy prediction based on energy consumption in a house, has the potential of not only regulating the utility consumption pattern but also, gives a predictive demand model for the house, which can be used in studying future demands. This paper examines the accuracy of different predictive models (KNN and SVM) based on U. Massachusetts's 2013 energy& occupancy datasets and predicts the occupancy based on 2015 energy consumption data using Machine Leaning algorithm.

KEYWORDS

Built Space Energy Consumption; Time-Variant Energy Pricing; Occupancy Prediction; KNN and SVM Models; Machine Learning

I. INTRODUCTION

Occupancy detection has been one of the main block of commercial and residential building automation systems. Efficient Heating, Ventilation and Air Conditioning (HVAC) systems rely on estimated occupancy data to control building's temperature and air flow [3]. Similarly, many lighting systems rely on the detection (presence or absence) of people in doorways or rooms to switch on and off [2]. Building occupancy detection is still a cumbersome and error-prone process despite a large number of intense research activity in studying the similar spectrum. Some of the challenges in occupancy prediction modeling comes from the data acquisition level. Most of the time the collected data are not a large data base. That is an effective system and appliance level energy consumption and occupancy data are not easily available. Moreover, it is a costly effort to install, and maintain the different sensors that can collect better and precise data at both the levels. Also, the logistical issues like power-cable sensors, availability of sensors, maintenance issues, faulty installations and non-expertise supervision of these collection sensors are major issues in data collection process.

In this paper we study the possibility of using electric meter data as a part of an opportunistic occupancy sensing infrastructure in domestic settings

II. DATASET DESCRIPTION

A. Introduction:

The metered electricity data and occupancy data is obtained from the Trace Repository, a database of The University of Massachusetts. The original energy consumption data consists of 5 sample houses (A, B, C, D, E) collected over a 3-year period (2013, 2014, 2015). However, for the scope of this paper dataset from houses A and B have been considered.

For training purpose, House B energy and occupancy dataset for the year 2013 has been used. This data is collected at one-minute interval resolution for a period of one week. It is to be noted, both Houses A and B have 2 occupancies each, so there are 3 cases for occupancy (0 = No Occupancy, 1= Only one person, 2= fully occupied). For the case of the simplicity, these experimental houses were not allowed any guests during the period of data collection. This simplifies the process of normalization when two building models are compared.

B. Data Exploration:

The 2013 datasets were used as a basis to train the model, so as to predict the occupancy from the 2015 energy dataset. This study provides a method to predict future occupancy based on current electricity consumption values.

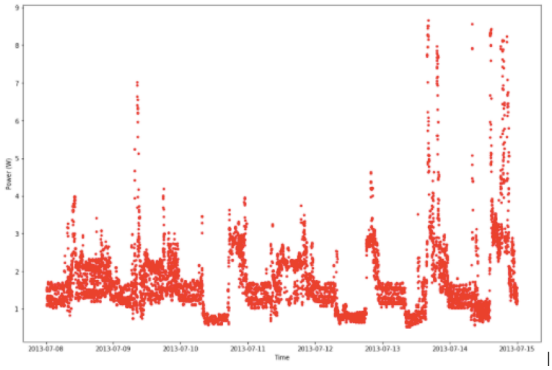

Figure 1: Power consumption data point of 2013 dataset

At first, the 2015 dataset is explored to identify the average hourly consumption as shown in figure [2](a) and (b). It can be seen that the 'Dining Room' and the 'Living Room' has high consumption during the evening-night time when there are occupants. Similarly, at appliance level, Appliances like Microwave, A.C., Dryer is mainly used during evening. However, it is to be noted that A.C. system and the furnace consume maximum energy.

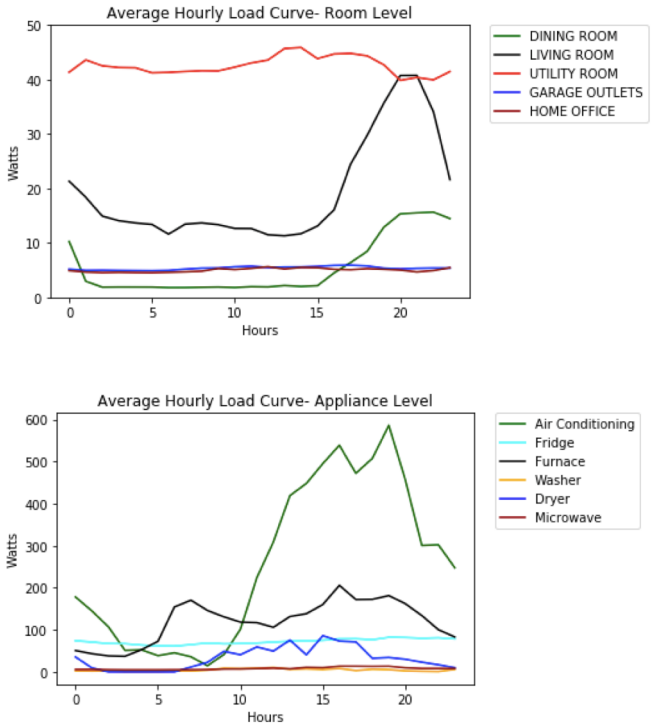

Figure 2(a): Average Hourly Load Curve – Room Lighting Level
Figure 2 (b): Average Hourly Load Curve – Appliance Level

C. Data Cleaning:

'Time Normalization' was carried out to match the 2013 occupancy and power consumption data. The final data set has one-minute time resolution for 1 week with 10,080 data points.

The 2015 energy data has continuous timestamp for 30-minute resolution for 1 year with 17,520 data points as shown in Figure (3).

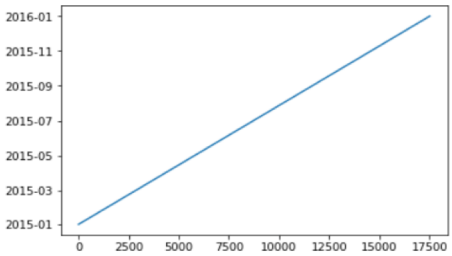

Figure 3: Power Data Timestamp for one year

III. METHODOLOGY

In this method we have performed stateless classifiers: (A) Support Vector Algorithm, (B) K-Nearest Neighbor (KNN)

The first method explored is a support vector machine (SVM) with a support vector classification as a supervised learning tool to analyze the 2013 Home A and B training data. This model assigns power consumption data to every occupancy data 'type' (0,1 or 2) and then builds a model based on trained dataset. In our model we have assigned 30% data points for training purpose with alpha value as 0.1 and 'C' as 1.

The second method explored is k-nearest neighbors (KNN), which classifies the object, e.g. – power use, by a majority vote of its neighbors. The power use is then assigned to the occupancy category most common among its "k" nearest neighbor. The algorithm finds the optimal 'k' or the number of neighbors. In our model, we have assigned 30% data points for training purpose.

The last method used is K-means clustering to train the model on sorting electricity demand into one of the three clusters (C_0, C_1, C_2) depending on its recorded occupancy. Then with the test set data, it can be stored into one of the defined clusters by calculating the difference in values in that cluster. Whichever cluster centroid has the smallest difference with means, it is classified as having occupancy as zero, one or two.

Also, we can even draw out their timestamps to plot a graph showing when the house is occupied and by how many people. This will give us a general trend whether the house is occupied or not. This information is particularly pertinent in balancing loads and predicting demands for the utilities.

IV. ALGORITHM AND ACCURACY

A. Algorithms:

The two classifiers used to predict the occupancy level are: (a) SVM and (b) KNN

SVM:

SVM is a supervised learning that associates learning algorithms that analyses data used for classification and regression analysis. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

$$\left[\frac{1}{n}\sum_{i=1}^{n}\max\left(0, 1 - y_i(w \cdot x_i - b)\right)\right] + \lambda\|w\|^2$$

Eq. [1]

As shown in the equation 1, The primary algorithm tries to maximize the distance of the gaps such the points are clearly distinguishable to separate zones.

KNN:

KNN is a type of instance based learning, where the function is only approximated locally and all computation is deferred until classification. A peculiarity of the KNN algorithm is that it is sensitive to the local structure of the data.

Euclidian Distance Measuring:

$$d_E(x, y) = \sum_{i=1}^{N}\sqrt{{x_i}^2 - {y_i}^2}$$

Eq. [2]

Scaling of resulting distances by arithmetic mean:

$$x' = \frac{x - \bar{x}}{\sigma(x)}$$

Eq. [3]

Arithmetic Mean and Standard Deviation:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Eq. [4]

$$\sigma(x) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Eq. [5]

B. Classification of Algorithms:

For KNN classifier we use *preprocessing, cross_validation* and *neighbors* from *sklearn* library. To implement SVM classifier we used *svm.SVC* function from scikit learn.

C. Classification Accuracy:

Table 1 shows the accuracy obtained from testing data on 2015 data set for the two houses. It can be seen that the accuracy in SVM for House A is very low.

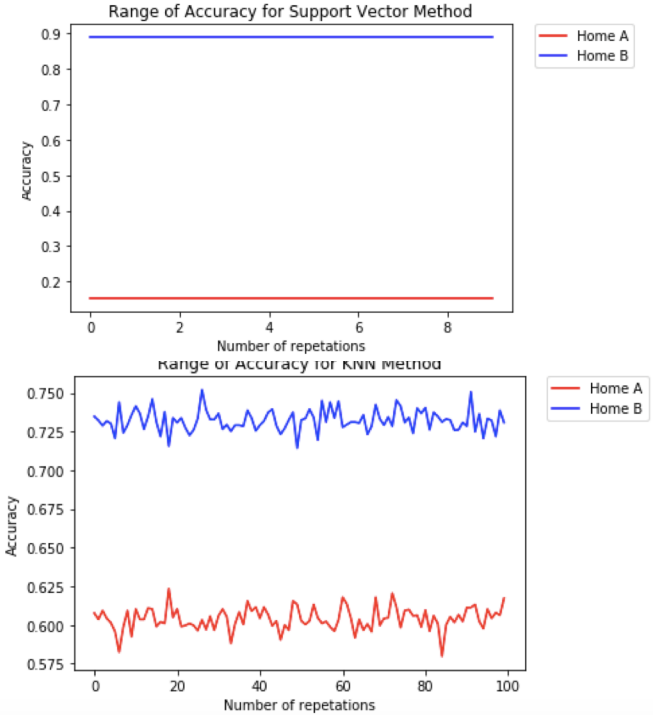

Figure 3(a): Range of accuracy for SVM
Figure 3 (b): Range of accuracy for KNN

Table 1: Accuracy of models for different houses

	House A	House B
SVM	15%	89%
KNN	60%	73%

From the table 1, we can infer that the prediction for house B is better in both the models compared to the prediction for house A. This shows that house A has more uncertainty in occupancy data compared to House B.

D. Results:

After training the model and checking for accuracy. This model is tested on the 2015 dataset to predict the occupancy based on the power consumption. Here, we do not have the 'ground truth' data set to compare the validity of the result. The 73% accuracy KNN model is applied to the 2015 data set for generating a prediction. Figure 4 shows the hourly frequency of event (C_0, C_1, C_2).

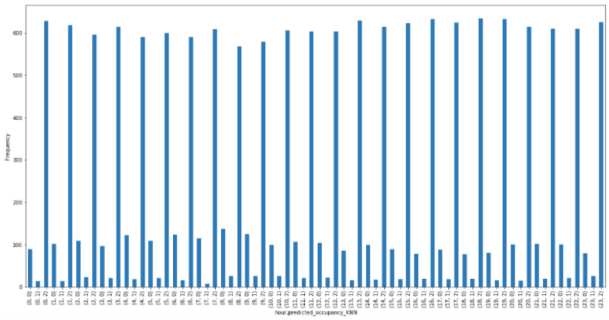

Figure 4: Hourly Frequency of Occupancy Events for 2015 House 'B'

From the K-means clustering we can find the hourly power consumption pattern of the house in 2015. In the K-means cluster model, average power consumption is grouped into three distinct time intervals (12:00 to 8:00, 8:00 to 16:00, 16:00 to 24:00). A general trend can be seen where power consumption peaks slightly around 8:00 to 9:00 and significantly around 15:00 to 16:00, with highest cluster average being greatest in the third time interval.

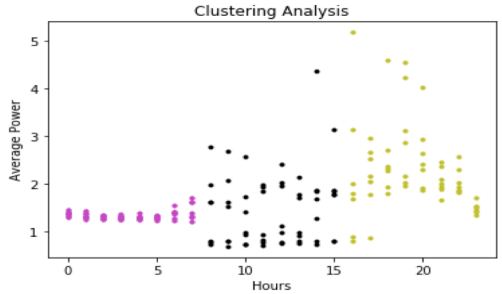

Figure 5: K-means clustering Analysis

V. CONCLUSION AND FUTURE WORK

The different models created here through a diverse array of methods, modest fits were achieved in predicting occupancy from electricity consumption. Each attempt from SVM to KNN to k-Means created models of different accuracy, but the relative scarcity of ground truth occupancy data hindered the ability to train the models well enough. However, given only 7 days of data to train and test on one-year dataset, the accuracy of 70% and 89% for house B results fairly enough.

As it exists now, this model can only be applied to small residential homes. In future, different models like HMM and THR can be applied to validate the results found by us. Also, a more bulk dataset with longer time duration shall help in achieving more accuracy making the analysis both more robust and applicable to a wider range of homes and locations.

ACKNOWLEDGMENTS

This work was partially done under the requirement for the coursework 12-752 Data Driven Building Energy management. A special thanks to Professor Mario Berges and Henning for their conceptual guidance on this subject.

REFERENCES

[1] W. Kleiminger, C. Beckel, T. Staake, S. Santini, 2013. *Occupancy Detection from Electricity Consumption*, ETH Zurich Published Paper, DOI: 10.1145/2528282.2528295.

[2] X. Guo, D.Tiller, G.Henze, 2010. *The performance of occupancy based lighting control systems: A review*. Lighting and Research Technology, DOI: 42(4): 415-431

[3] C.Beckel, L.Sadamori, S. Santini, 2013. *Automatic socio-economic classification of households using electricity consumption data.* Proc e-Energy, ACM