The User Side of Sustainability: Modeling Behavior and Energy Usage in the Home

Chao Chen[a,*], Diane J. Cook[a], Aaron S. Crandall[a]

[a]*School of Electrical Engineering and Computer Science, Washington State University, Pullman, WA, USA*

Abstract

Society is becoming increasingly aware of the impact that our lifestyle choices make on energy usage and the environment. As a result, research attention is being directed toward green technology, environmentally-friendly building designs, and smart grids. This paper looks at the user side of sustainability. In particular, it looks at energy consumption in everyday home environments to examine the relationship between behavioral patterns and energy consumption. It first demonstrates how data mining techniques may be used to find patterns and anomalies in smart home-based energy data. Next, it describes a method to correlate home-based activities with electricity usage. Finally, it describes how this information could inform users about their personal energy consumption and to support activities in a more energy-efficient manner. These approaches are validated by using real energy data collected in a set of smart home testbeds.

Keywords: smart environments, machine learning, energy, anomaly detection

1. Introduction

In 2010, the United States consumed 98,003 Quadrillion Btu of power energy. This is a 200 percent increase from 1949 [1]. The growth of energy usage is not entirely due to manufacturing plants and automobiles, as is often assumed. In fact, worldwide residential sector is responsible for 16-50% of energy consumption consumed by all sectors [2]. As a result, there is an urgent need to develop

*Corresponding author

Email addresses: cchen@eecs.wsu.edu (Chao Chen), cook@eecs.wsu.edu (Diane J. Cook), acrandal@wsu.edu (Aaron S. Crandall)

technologies that examine energy usage in homes and to encourage energy efficient behaviors, in addition to energy efficient devices in households.

Although households and buildings are responsible for over 40% of energy usage in most countries [3], many residents still receive little or no detailed feedback about their personal energy usage. A power utility bill traditionally provides information about a month's total energy consumption and a total price to be paid, leaving homeowners to guess what factors, including external influences and internal behavior, might explain a higher or lower than usual bill. Earlier studies have shown that home residents reduce energy expenditure by 5-15% just as a response to acquiring and viewing raw energy usage [4]. Residential behavior, which varies widely, can influence energy usage significantly in a given home [5]. Clearly, the typical utility bill provides no information about the relationship between residential behavior and corresponding energy usage. Since behavior-based energy information is capable of encouraging individuals to modify habits in ways that would be beneficial for both the household and community, it would be desirable to develop technologies that could extract information from the home and communicate it to residents. However, occupants' behavior is difficult to capture accurately. Self-report of behavior is error prone [6] and whole-home meter monitoring does not capture the behaviors in the home that influence consumption.

We hypothesize that providing users with knowledge about the relationship between their activities and energy consumption, suggestions for energy reduction, and automation support will result in more substantial decreases in overall consumption. This view is supported by an increasing body of work that links awareness of energy consumption and its impact to behavioral change [7, 8]. In our work we propose using smart homes and pervasive computing techniques to provide these important insights. The long-term vision for this project is to enhance understanding of human resource consumption and to provide resource efficiency in smart homes. We envision this as a three-step process: 1) analyze electricity usage to identify clusters and anomalies, 2) correlate activities with energy usage, and 3) automate energy-efficient activity support. Additionally, we hypothesize that patterns and anomalies may be automatically detected from energy consumption data and that these discoveries can provide insights on behavioral patterns. We further postulate that energy consumption is correlated with the type of activities that are performed and can therefore be predicted based on the automatically-recognized activities that occur in a smart environment. These hypotheses are validated by implementing algorithms to perform these steps and evaluating the algorithms using data collected in the smart apartment testbeds. Finally, a discussion of how the results of this work can be used to give smart home

residents feedback on their energy consumption is included. This work represents one of the first projects that utilizes smart home data to investigate the relationship between behavioral patterns and resource consumption in a home environment.

2. Related Work

A smart home environment can be defined as one that acquires and applies knowledge about its residents and their physical surroundings in order to improve their experience in that setting [9]. Such home environments, equipped with sensors for detecting features such as motion, light level, temperature, and energy and water consumption, are ideal testbeds for investigating the relationship between behavior and energy consumption. Using sensor technology combined with data mining and machine learning, many researchers are now working on smart environments, which can discover and recognize residents activities and respond to their needs in a context-aware way. As household consumption of electricity has been growing dramatically, the need to develop technologies that improve energy efficiency and monitor energy usage in a household is emerging as a critical research area.

Technologies to address this need are beginning to emerge. Non-intrusive appliance load monitoring [10] has been designed to detect the turning on and off of individual appliances in an electrical circuit. Several academic studies focused on this topic to estimate residential energy levels based on appliance usage [11, 12]. With respect to energy conservation, some industrial products focus on providing energy information services and saving tips to residents. Google PowerMeter [13] is a free energy monitoring tool for saving energy by providing energy information via smart meters. The companies, such as Microsoft Hohm [14] and Opower [15], apply statistical methods and data mining algorithms to analyze raw utility data and give customers usable energy saving tips. However, these projects are orthogonal to this paper, in which we provide users with feedback that is actually related to resident behavior in the home. Likewise, several studies exist that predict building energy consumption at a highly aggregated level for a large collection of buildings [16], but these studies also differ from our work, as we consider human behaviors in an individual building as primary features for predicting energy usage.

Anomaly detection finds extensive use in a wide variety of applications such as intrusion detection fault detection for credit cards, medical health, and sensor networks. Since society is becoming increasingly aware of the impact of energy efficiency, anomaly detection has been directed toward energy aspects of building

management. Grwtham et al. [17] translate time series data to frequency spectrum, and then use a K-NN algorithm to identify anomalies in sparse regions. This method can detect abnormal patterns with low frequency, but ignores more common patterns with anomalously high fluctuation. Other efforts [18, 19] first extract the features from daily energy consumption then use statistical methods to identify abnormally high or low energy use. However, these methods relied on the assumption that the data is sampled from a particular distribution, which may not hold true. It can also be difficult to identify contextual anomalies with small fluctuations.

In comparison to other past works, several different contributions are offered by this paper. The data sets used in this work are capable of representing realistic residential patterns over the monitoring period, since this monitoring did not affect the residents' daily routine. Machine learning techniques are evaluated to explore the relationship between residential activities and energy usage and build predictive models. For anomaly detection, this approach transforms raw energy data into a symbol sequence, and then extends a suffix-tree data structure to analyze structural patterns. Specific metrics are expected to detect three various types of anomalies: 1) anomalies diverge notably from other patterns; 2) anomalies whose inner has large fluctuation; 3) anomalies that occur occasionally. To the best of our knowledge this is the first work that applies pattern-discovery to detecting energy outliers in home environments.

3. CASAS-Sustain System Architecture

In this paper, we describe a prototype system framework for energy data collection, energy data transformation, and energy data analysis, as shown in Figure 1. The system, called CASAS-Sustain, operates entirely within the structure of the CASAS smart home project [20]. As the diagram indicates, data collected in the smart home is first analyzed to look for patterns and outliers. Next, recognized activities in the home are correlated with energy usage to provide working information on the energy usage that is usually required to support a class of activities. Finally, the smart home can suggest or automate control of devices that are not required during the current activity in order to reduce energy expenditure and wasted resources.

3.1. The Smart Home Environment

The smart home environment testbeds used to analyze energy usage are two apartments located on the Washington State University campus. As shown in Fig-

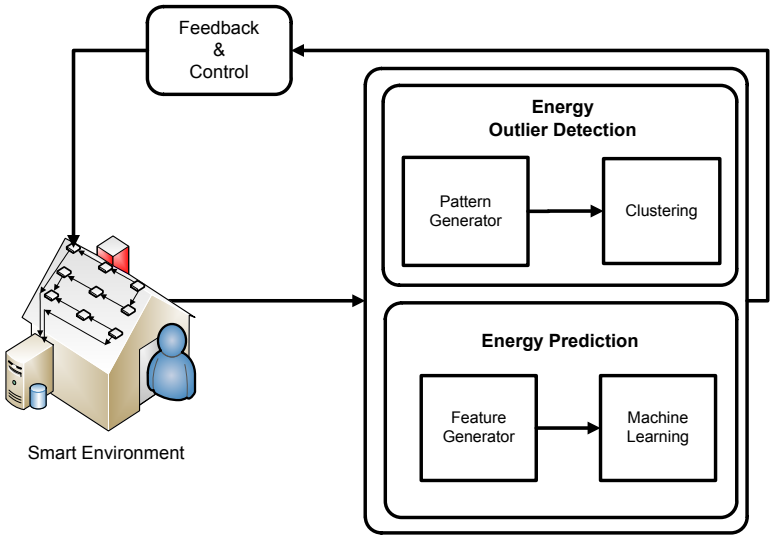


Figure 1: Architecture of the CASAS-Sustain system.

ure 2, the Kyoto smart home apartment testbed consists of three bedrooms, one bathroom, a kitchen, and a living/dining room. We also monitor the two-floor Tulum apartment, which consists of two bedrroms, a living room, dining room, kitchen, and bathroom. To track people's mobility, we use motion sensors placed on the ceilings. The circles in the figure stand for the positions of motion sensors. They facilitate tracking the residents who are moving through the space. In addition, the testbed also includes temperature sensors as well as custom-built analog sensors to provide temperature readings and hot water, cold water and stove burner use. A power meter records the amount of instantaneous power usage and the total amount of power, which is used. An in-house sensor network captures all sensor events and stores them in a SQL database for long term storage. Our research team has installed 26 of these types of smart homes around the Pacific Northwest at a cost of approximately $3,000 per home. Installation takes 2-3 hours and removing the smart home equipment takes about 30 minutes. The components are not intrusive and the residents often forget they are present after the first week. As a result, using this technology is a fairly realistic approach to providing context-aware services and investigating the link between behavior and sustainability.

The sensor data gathered for our SQL database is expressed by several features, summarized in Table 1. These four fields (Date, Time, Sensor ID and Message) are generated by the CASAS data collection system automatically. To provide real training data, data was collected from both of the smart apartments

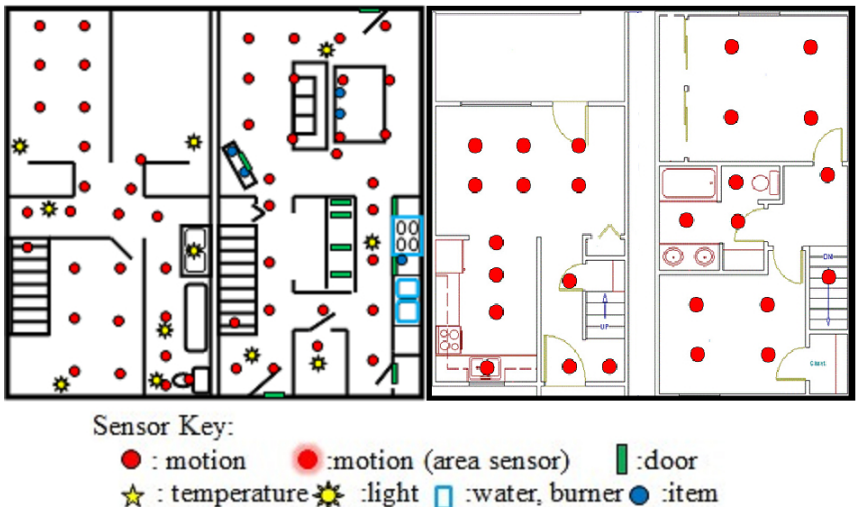


Figure 2: CASAS smart apartment testbeds: Kyoto (left) and Tulum (right).

Table 1: Raw data from sensors. Sensor IDs beginning with "M" refer to motion sensors, "T" refers to temperature sensors, and "P" refers to power readings.

Date	Time	Senosr ID	Message
2009-02-06	17:17:36	M45	ON
2009-02-06	17:17:40	M45	OFF
2009-02-06	11:13:26	T004	21.5
2009-02-06	11:18:37	P001	930W

included in this study while two pairs of adult volunteers in good health were living in the smart apartments. Our training data was gathered over several months and more than 100,000 sensor events were generated in each site during this time. All of our experimental data are produced by the day to day lives of these residents, which guarantee that the results of this analysis are reflective of their routine behavior.

4. Energy Data Analysis Outlier Detection

Our first step in utilizing smart home technologies for energy efficiency is to better understand the nature of the energy consumption itself. We begin by analyzing normal patterns of usage and identifying abnormal or anomalous situations. We analyze normal patterns by clustering sequences of power usage values. This analysis is useful because the cluster descriptions can provide users with insights

on their daily habits and resource usage as well as provide software algorithms with a model of normal usage in a particular environment. At the same time, the clusters provide a baseline against which anomalies in energy usage may be identified. Anomaly detection is valuable because the anomaly may indicate an unnecessary use of resources (e.g., an appliance was accidentally left on), an unsafe state, or noise in the dataset. We begin with definitions for the data we are analyzing.

Definition 1. *Let $e = (t, v)$ be an individual **energy senosr event** in our smart environment, where t refers to the timestamp when v has been generated, and v refers to an energy numeric value.*

For data mining purposes, we are typically not interested in any individual energy sensor event. Rather, we are interested in a sequence of these energy values:

Definition 2. *An **energy sensor sequence** $E = e_1 e_2 \dots e_n$ is a time-ordered set of n energy sensor events.*

As is shown in Table 1, smart home power meters record the amount of instantaneous power that is currently being consumed in real time. To search the pattern more efficiently, we first discretize this data into k value ranges using equal-width binning [21] and then convert the value ranges to symbols. This representation allows symbolic approaches to be applied to analyzing the data, at the risk of losing some precision in the values. Another advantage is that it allows the use of data structures and algorithms that are not well defined for real-valued data, including the suffix tree structure we used in our research. Through binning, an energy sensor sequence E can be transformed into a discretized energy symbolic sequence S, which is defined as:

Definition 3. *An **energy symbol sequence** $S = s_1 s_2 \dots s_n$ is an ordered set of n symbol variables over the alphabet Σ, where $\Sigma = \{a, b, c, \dots\}$ and $\|\Sigma\|$ is equal to the number of bins k. All energy values in the range for the i^{th} bin are represented by symbol i in the sequence.*

After converting the raw power data into a symbolic sequence, our algorithm employs a suffix tree [22] to discover sequential, recurring patterns of energy usage. Unlike other data mining methods, which are exponential in their complexity, this approach can generate a suffix tree in $O(n)$ time for a symbol sequence of length n, and spend $O(m)$ time searching for a subsequence of length m, regardless of n. A formal definition of this tree follows.

Definition 4. *Given a string $S^{'}$ over the alphabet Σ and a unique termination character $\$ \notin \Sigma$, the string resulting from appending $\$$ to $S^{'}$ can be defined as $S = S^{'}\$$. Let $|S| = n$ and $suf f_{S,i} = S_i S_{i+1} \dots S_{|S|}$ be the suffix of the string S starting at i^{th} position. The **suffix tree** of S is a compacted trie-like data structure that stores all suffixes of a string S over the alphabet Σ.*

Traditional suffix tree construction algorithms start from the root and follow a unique path matching characters in $suf f_{S,i}$ one by one until no more matches are possible. If the traversal does not end at an internal node, it creates a new internal node at that location. For a tree with n nodes, the total running time of the algorithm is $\sum_{i=1}^{n}(n - i + 1) = O(n^2)$. In order to achieve $O(n)$ running time, we use McCreight's algorithm [23] to construct a suffix tree by applying suffix links to speed up the insertion of a new suffix.

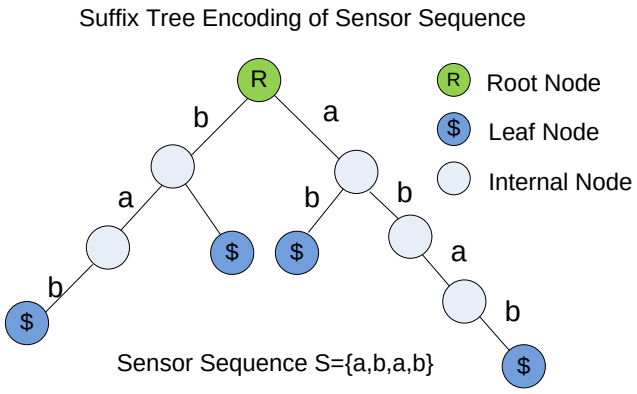


Figure 3: A suffix tree defined on a symbol sequence S with length m can represent every subsequence in S with at most $2m$ nodes

A graphical illustration of the transformation of an energy sequence into its equivalent suffix tree is shown in Figure 3. By definition, no two edges emanating from a node in a suffix tree begin with the same symbol, which implies that every unique subsequence in S starting from the root node can be generated by traversing through the suffix tree. We consider these subsequences as **energy patterns**, which are defined as:

Definition 5. *Let an **energy pattern** $p_i \in S$ represent the subsequence generated by traversing a path in the suffix tree, where p represents the sequence of symbols visited along the path and the length of this energy pattern is i. The frequency of*

an energy pattern $p_i \in S$ is denoted by $f(p_i)$, which is equal to the number of the leaf nodes found in the subtree rooted at the end of the subsequence p_i.

Table 2 shows two examples of energy patterns and their corresponding frequencies. In the first case, energy readings of 752 and 742 fall in the same bin (value range) and are mapped to symbol C. The sequence of energy readings CC occurs 26,592 times in the data file and thus is a much more common pattern than the one found in the second line of the table. In the context of this brief example sequence CC might be considered a pattern of interest, while sequence ZFZ might be considered an outlier or anomaly.

Table 2: Examples of Energy Pattern

Energy Pattern	**Pattern Length**	**Raw Energy (Watt)**	**Pattern Frequency**
CC	2	752 742	26952
ZFZ	3	5000 1021.2 5007	13

4.1. Sequence Clustering

To detect abnormal situations, we next cluster all the energy patterns into groups with similar patterns and identify sequences that do not fit well in any cluster. Intuitively, for an energy symbol sequence S, we consider an energy pattern p_i to be an outlier, if this energy pattern is far from the centroid of the cluster.

Cluster analysis is a data mining technique that is often used to identify various groupings or taxonomies in datasets. We apply clustering to power sequence values in order to gain a better understanding of the data, to identify groupings of normal energy usage, and to use as a baseline for identifying abnormal energy usage patterns. A clustering algorithm takes features of the data as input and creates a classification scheme which is represented as a set of disjoint clusters, each of which can be described by a middle point, or cluster centroid.

One important step in our clustering process is to decide a distance measure, which is used to group sequences together in a cluster and should reflect the similarity of two sequences. In this paper, we use a two-step process. We first restrict clusters to contain only patterns of the same length. That is because the suffix tree algorithm naturally groups patterns into distinct lengths, and our algorithm

further divides these groups into subgroups using the clustering algorithm. From these groups we next employ Euclidean distance measure, which is a geometric distance in the multidimensional space and is widely used by clustering algorithms. Based on specific property of energy patterns, we select three related features, which will be used to measure the dissimilarity, or distance between energy patterns.

Pattern Variance between Energy Patterns. This metric is mainly used to detect abnormal patterns. As defined in Definition 5, $p_i = s_1 s_2 \dots s_i$ is an energy pattern, where s is a discretized symbolic energy usage sequence. The distance between two symbols $|s_x - s_y|$ can be be estimated as the alpha-numeric distance between the symbols. To determine pattern variance, we measure the distance between each corresponding symbol in the pattern. Thus the pattern variance between $p^1$and p^2 with length i is defined as:

$$d_1(p^1, p^2) = \sum_{j=1}^{i} |s_j^1 - s_j^2| \tag{1}$$

Within-Pattern Variance. Because changes in power occur when appliances are switched on or off, the difference between two consecutive symbols in an energy pattern may indicate a change in the status of the appliances. Thus, the variance within this energy pattern can capture the usage status of the appliances. The within-pattern variance of an energy pattern p can be calculated as $v_i = \sum_{j=2}^{i} |s_j - s_{j-1}|$. We define the difference in within-pattern variance between two energy patterns p^1 and p^2 as:

$$d_2(p^1, p^2) = |v^1 - v^2| \tag{2}$$

Frequency of Energy Pattern. Another important feature we cannot ignore is the frequency of an energy pattern, as defined in Definition 5. The lower the frequency is, the more likely this pattern is an outlier. If the frequency of a pattern is relatively high, it may represent a normal pattern of usage. Therefore, this measurement is able to find out the patterns that occur rarely. The frequency difference between energy patterns p^1 and p^2 is calculated as:

$$d_3(p^1, p^2) = |f(p^1) - f(p^2)| \tag{3}$$

To balance the impact of these three metrics, all these three distance values are normalized to the scale [0, 1] and the final distance between two energy patterns p^1 and p^2 is estimated as:

$$d(p^1, p^2) = \sqrt{d_1(p^1, p^2)^2 + d_2(p^1, p^2)^2 + d_3(p^1, p^2)^2} \quad (4)$$

4.2. Outlier Detection

In the second step of our analysis, we use the generated clusters to identify outliers in the energy usage data. The outliers are defined as energy usage sequences that fall as far as possible from the centroid of any cluster. Detecting these outliers consists of two stages. In the first stage, we cluster the energy sequences and calculate the cluster centroids. In the second stage, we calculate the distance of each energy pattern sequence to the cluster centroids. The greater this distance is, the more likely it is that the pattern is an outlier. Patterns for which the distance is greater than a pre-defined threshold are considered to be outliers and indicate anomalous energy usage.

From this discussion it is apparent that the choice of a threshold value greatly influences the selection of outliers. A variety of methods could be used to select the threshold. These could be based on statistical parameters of the data itself or user-selected parameters such as the rarity of the anomalies that are being reported. To determine the value for this application domain we plot a histogram of all pattern distance values to the centroid (also referred to as outlying factors, see Figure 4. It was noted that these outlying factors follow a normal distribution, which means that 99.7% of the patterns will then fall within three standard deviations of the mean. To detect the outliers, we only consider the patterns that fall outside of this area.

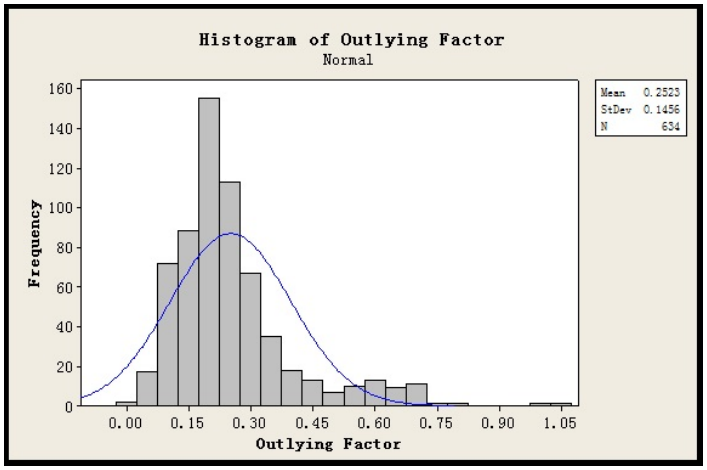


Figure 4: Histogram of outlying factors of all energy patterns ($k = 50$ bins).

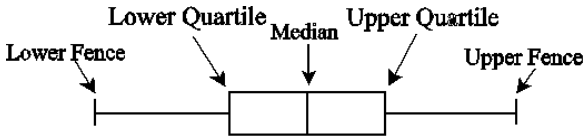


Figure 5: Configuration of a box plot.

In our study, we use the box plot [24] as an alternative method to identify outliers in the collected energy data, which represent those periods of time where the energy consumption lies unusually far from the main body of the data. If Q_1 and Q_3 are the lower quartile and the upper quartile, a measure of spread that is resistant to the outliers is the inter-quartile range or IQ, calculated as $IQ = Q_3 - Q_1$. As shown in Figure 5, the fences lie at $Q_1 - k * IQ$ and $Q_3 + k * IQ$. The change of the value of k can affect the number of the observations outside the fence. For this work, a value of $k = 1.5$ was used, which has been indicated as acceptable for most situations. Any sample data farther than $1.5 * IQ$ from the closest quartile is an outlier. An outlier is extreme if it is more than $3 * IQ$ from the nearest quartile and it is mild otherwise.

4.3. Experimental Results

Two series of experiments were performed using energy data collected from two of our smart apartments, which we name Kyoto and Tulum. In the first group of experiments, the standard boxplot method to detect outliers in two months of energy data from the Kyoto testbed was used. In comparison, the proposed cluster-based approach was applied to detect the outliers existing in the same dataset. The second group of experiments looks for abnormal energy data during a single week from both Kyoto and Tulum, respectively.

For the first experiment, a total of 95,968 power events were collected. Figure 6 shows the result of the boxplot approach on this dataset. The black points located on the top represent the outliers. The boxplot considered 12,718 sensor events as potential outliers, since it is merely able to detect energy consumption that lies unusually far from the main body of the data. However, it is difficult for users to determine which outliers are true outliers and identify potential reasons for these outliers, because there are too many false positives. Moreover, statistical methods also cannot identify abnormal energy patterns our technique can detect.

Next, we use our proposed clustering algorithm to analyze the same power dataset. Figure 7 depicts the distribution of energy patterns that were detected as potential outliers for alternative numbers of bins and clusters. It should be noted

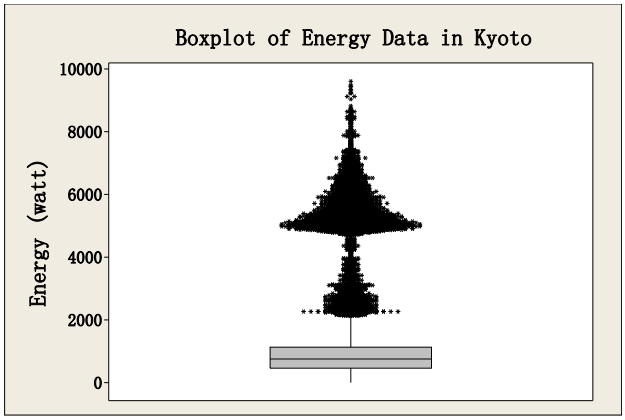


Figure 6: Energy outlier detection using the boxplot.

that the clusters generated by all the patterns with various lengths are marked as 'Pattern Length' in the figure. Comparing our method with the boxplot, it shows that the number of the outliers reported by the clustering approach has been decreased notably. This increases the chance to accurately determine real outliers in the dataset.

Table 3: Experimental results of outlier detection (k=30 bins)

Kyoto		Tulum	
Pattern Length	**Number of Outliers**	**Pattern Length**	**Number of Outliers**
2	0	2	0
3	3	3	10
4	5	4	21
5	4	5	18
6	13	6	13

The second group of experiments focuses on energy usage data collected for a single week in the Kyoto and Tulum testbeds. The purpose of this experiment was to detect energy outliers and determine possible reasons for these outliers. Table 3 displays the results of the clustering method for 30 bins. To explore potential reasons for the anomalous usage patterns, those outliers were examined in detail. It was discovered that these abnormal events represent two types of occurrences. The first set of outliers was mainly due to large changes in energy usage, when the residents had sustained high-level energy consumption over a long time. Some of

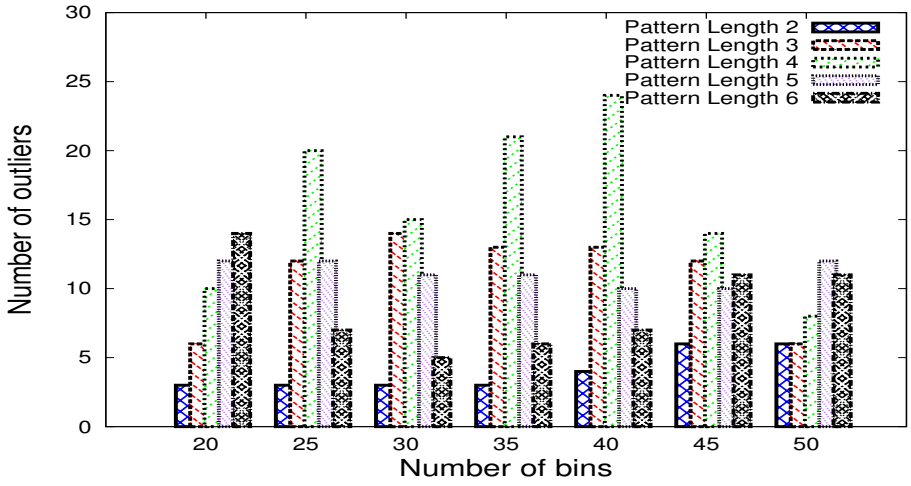


Figure 7: Distribution of number of pattern outliers using our clustering approach.

the big appliances, including the water heater, consume more energy than others and can create anomalies when there are long showers. In addition, during the middle of the day the residents often do their cooking and large appliances are being used for cooking such as the microwave, the stove and the oven, all of which would give rise to a dramatic increase in energy consumption. To respond to these outliers, the residents can analyze their energy needs during these activities to identify energy-saving behaviors.

Table 4: Example of an Outlier.

2009-06-01 23:31:02	P001	1001.5
2009-06-01 23:31:02	P001	356

The outliers in the second set consisted of two successive energy events, whose values are different but occur at the same time, as shown in Table 4. This situation actually represents noise in the data that occurs as part of the data collection hardware. These kinds of outliers are also valuable to detect because the noise can be addressed to subsequently improve the accuracy of additional analysis methods. Therefore, we checked the entire Kyoto and Tulum datasets for these types of outliers. The result was that 6,398 entries from Kyoto and 9,401 entries from Tulum that represented noisy data collection conditions were removed. In the second group of experiments, all of the outliers detected by the clustering approach fit into one of these two categories. However, since we only consider the patterns,

which are extremely far from the centroid of the clustering, the rate of false negatives may be somewhat higher, which means that some real outliers are likely to be ignored by this approach. One possible solution is to decrease the pre-defined threshold, which makes our approach to detect more outliers with the risk of increasing the rate of false positive.

5. Activity-based Energy Prediction

In the second step of our CASAS-Sustain analysis, machine-learning techniques were used to predict energy consumption given information about an activity that residents perform in a smart environment. Because activity recognition techniques are prevalent in the literature [25] and are becoming more robust, this offers a practical approach to automatically correlating activities in the home with energy consumption. The following features were used to describe an activity performed by an inhabitant in a smart home:

Table 5: Data features for classification models.

Feature Name	**Description**
Activity label	This feature indicates the types of the activities the residents perform.
Activity length (in seconds)	This feature shows the time length of the activity.
Day of week	This feature shows the current day of week (Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, and Sunday).
Weekday/Weekend	This feature is a binary variable to determine whether the current day is a weekday or weekend.
Time of day	This feature represents different time slots (morning, noon, afternoon, evening, night, and late night).
Times of individual sensors triggered	This feature represents times of different motion sensors that were activated during the activity
Number of motion sensors activated in various rooms	This feature represents the number of motion sensors that were triggered in various rooms.
Total number of motion sensor events triggered	This feature represents the total number of motion sensor events that were triggered

The features were used to describe an activity performed by an inhabitant in a smart home as shown in Table 5. The input to the learning algorithm is a list of these eight features as computed for a particular activity that was performed.

The output of the learning algorithm is the amount of electricity that is predicted to be consumed while performing the activity. In this paper, by applying equal-width binning, the target average energy data was discretized into several interval sizes (two classes, three classes, four classes, five classes, six classes, and seven classes) to assess the performance of our experiments.

Table 6: Electrical appliances associated with each activity.

Activity	**Appliances Directly Associated**	**Associated Appliances**
Work at computer	Computer, printer	Localized lights
Sleep	None	None
Cook	Microwave, oven, stove	Kitchen lights
Watch TV	TV, DVD player	Localized lights
Shower	Water heater	Localized lights
Eating	TV	Localized lights

To train the algorithms, sensor events were annotated with the corresponding activities being performed while the sensor events were generated. All of the activities that the participants perform have some relationship with measurable features such as the time of day, the participants's movement patterns throughout the space, and the on/off status of various electrical appliances. These activities are either directly or indirectly associated with a number of electrical appliances and therefore have a unique pattern of power consumption. Table 6 lists the appliances that are associated with each activity. It should be noted that, there are some appliances which are always on, such as the heater (in winter), refrigerator, phone charger, etc. Thus, we postulate that the activities will have a measurable relationship with the energy usage of these appliances as well.

5.1. Analysis of Resident Activities and Energy Usage

Figure 8 shows the energy fluctuation that occurred during a single day on June 2nd, 2009. The activities are indicated by the arrows. The length of the arrows indicates the duration of time (not to scale) for the activities. Note that there are a number of peaks in the graph even though these peaks do not always directly correspond to a known activity. These peaks are due to the water heater, which has the highest energy consumption among all appliances, even though it was not used directly. The water heater starts heating by itself whenever the temperature of water falls below a certain threshold.

Figure 9 plots typical energy data for each activity together with the result of applying curve fitting to the data. Curve fitting [26] is the process of building a

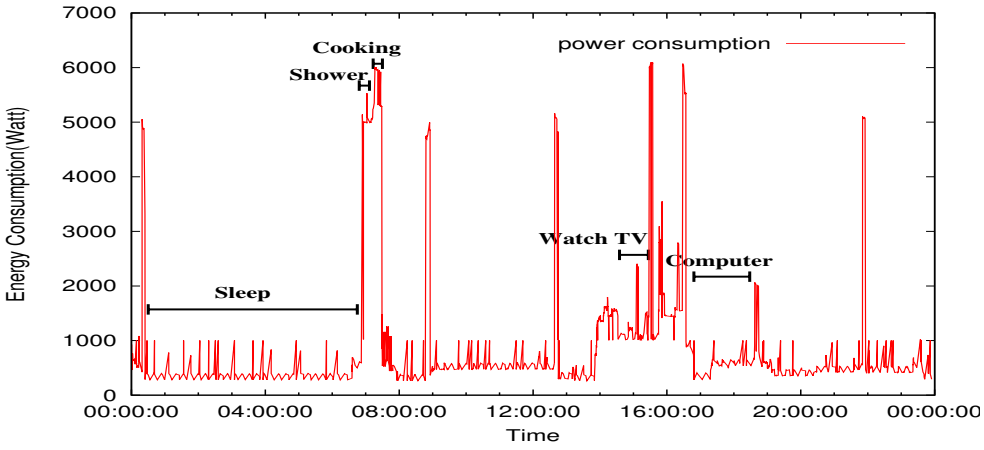


Figure 8: Energy usage for a single day.

mathematical function model that can best fit to a series of data points. It serves as an aid for data visualization, and to express the relationships between different data points. From the figure, we see that each activity generates a different energy pattern.

Figure 10 illustrates two boxplot graphs of energy consumption for each activity for both Kyoto and Tulum testbeds. From the graph, it again shows that each activity utilizes very different amounts of power. In Kyoto, the shower activity consumes the highest amount of energy because the water heater is a larger power consumer. However, the cook activity uses the most energy in Tulum. Cooking in Tulum involves frequent access to the refrigerator and stove, which increases power consumption. Conversely, when the participants were sleeping in both two testbeds the energy consumption was the lowest because most appliances were idle.

5.2. Modeling of Activity-Based Energy Usage

Machine learning algorithms are capable of learning and recognizing complex patterns contained in sensor data. In this work, machine learning algorithms were used to map these activity features onto a class label indicating the amount of energy that is consumed in the smart environment while the activity was performed. Three popular machine-learning methods were leveraged and compared for this work: a Bayesian belief network classifier, a support vector machine, and a neural network.

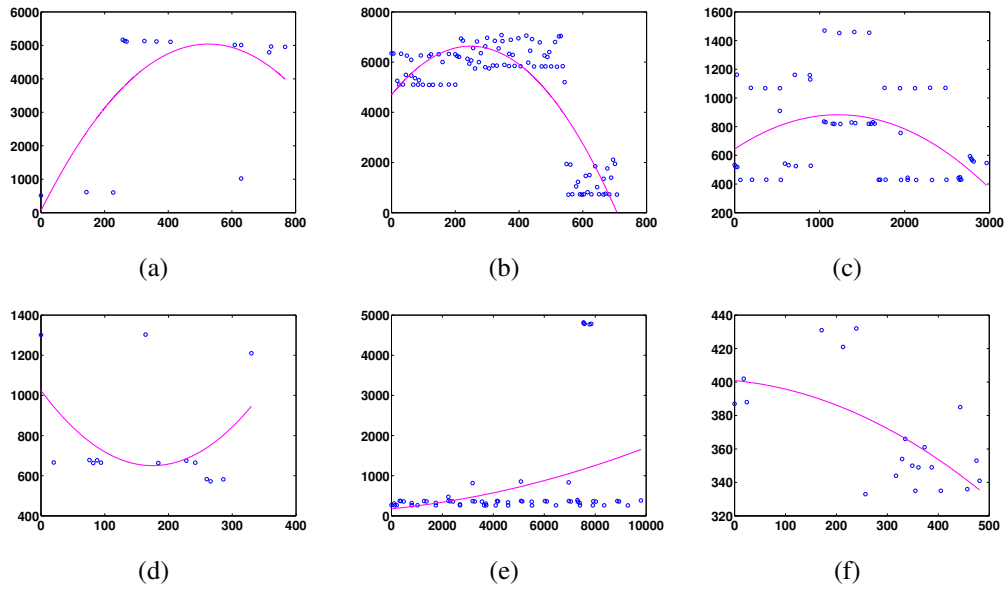
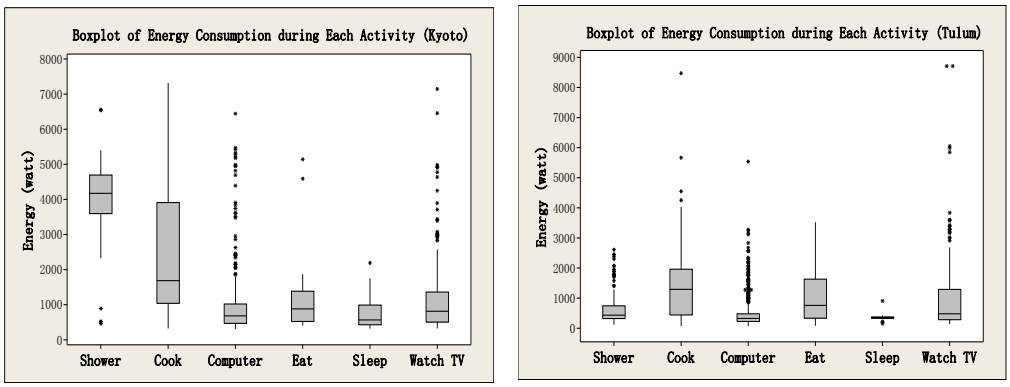


Figure 9: Energy data curve fitting for each activity. There is a separate graph for each activity: a=shower, b=cook, c=work on computer, d=eat, e=sleep, and f=watch TV. The x-axis in the graphs represents wattage and the y-axis represents time of the activity in seconds.

Bayesian belief networks (BBNs)[27] represent a set of conditional independence assumptions by a directed acyclic graph, whose nodes represent random variables and edges represent direct dependence among the variables and are drawn by arrows by the variable name. Unlike the Naïve Bayes classifier, which assumes that the values of all the attributes are conditionally independent given the target value, Bayesian belief networks apply conditional independence assumptions only to subsets of the variables. They can be suitable for small and incomplete data sets and they incorporate knowledge from different sources. After the model is built, they can also provide fast responses to queries.

Support Vector Machines (SVMs) [28] are a class of training algorithms for data classification, which maximize the margin between the training examples and the class boundary. A SVM learns a hyper-plane which separates instances from multiple energy usage classes with maximum margin.

Artificial Neural Networks (ANNs) [29] are abstract computational models based on the organizational structure of the human brain. The most common learning method for ANNs, called Backpropagation, which performs a gradient descent within the solutions vector space to attempt to minimize the squared error



(a) (b)

Figure 10: Boxplot of energy data generated by human activates in the Kyoto (a) and Tulum (b) testbeds.

between the network output values and the target values for these outputs. Although there is no guarantee that an ANN will find the global minimum and the learning procedure may be quite slow, ANNs can be applied to problems where the relationships are dynamic or non-linear and capture many kinds of relationships that may be difficult to model by other machine learning methods. In our experiment, the Multilayer-Perceptron algorithm with Backpropagation to predict electricity usage was leveraged.

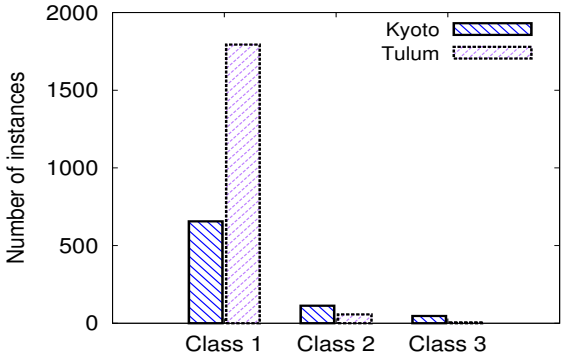


Figure 11: Distribution of instances in the three energy classes for Kyoto and Tulum.

One challenge that we face when learning a mapping from activity features to energy usage is that the class distribution is highly skewed, as is shown in Figure 11. This is because most home-based activities require a moderate amount

of energy usage, while a small set of activities require substantially more power. Machine learning algorithms are often challenge by such an imbalanced class distribution [30] because models that map all (or most) of the cases to the low-energy values will achieve high accuracy, but clearly will not learn the true mapping of activity features to energy usage.

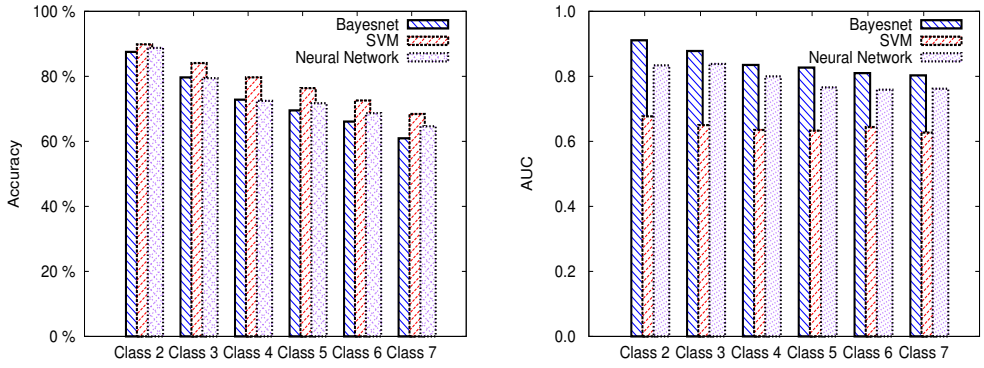
To deal with this imbalanced data we incorporate a data sampling technique, called SMOTE [31]. We combine under-sampling methods (to reduce data points in over-represented classes) with over-sampling methods (synthetically generating points for under-represented classes) to address this problem by using a combination of both under and over sampling, but without data replication. Here, over-sampling is performed by synthesizing a new sample corresponding to each minority class by randomly choosing from the points nearest neighbors. Generation of the synthetic sample is accomplished by first computing the difference between the feature vector (sample) under consideration and its nearest neighbor. Next, this difference is multiplied by a random number between 0 and 1. Finally, the product is added to the feature vector under consideration. The result is a new sample similar to, but not a replica of, the existing data.

5.3. Experimental Results

Two series of energy prediction experiments were performed. The first experiment uses the sensor data collected during two months in the Kyoto testbed. In the second experiment, we collected data of two months in the Tulum testbed. Using the Weka machine learning toolset [32], we assessed the classification accuracy of our three selected machine learning algorithms and reported the predictive accuracy results based on a 3-fold cross validation. It should be noted that the instances of the class in each group follow the real distribution to examine the performance of the sampling technique.

Conventional performance measures consider different classification errors as equally important. However, this assumption is not practical for our energy prediction, where the class distribution is highly skewed. Therefore, we consider two metrics that measure different aspects of performance. The first metric we use evaluates conventional accuracy of the classifiers; the second measurement is the area under a ROC curve (AUC), which evaluates overall classifier performance without taking into account class distribution or error cost.

Figures 12 and 13 plot the accuracies and AUC values for two set of experiments. The accuracy peaks around 90% for both datasets when predicting the two-class energy usage and the lowest accuracy is around of 70% for the seven-class



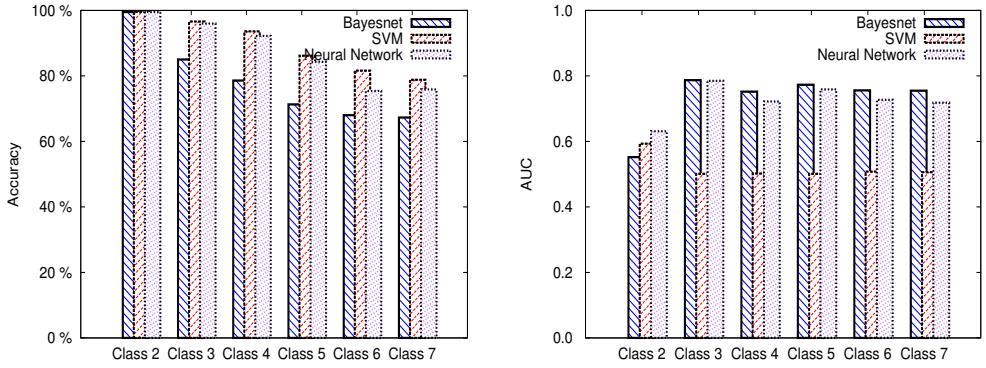
(a) Accuracy (b) AUC

Figure 12: Comparison of the accuracy and AUC for the Kyoto dataset .

case in both datasets. These results also show that the higher accuracy will be attained with a lower precision. Increasing the precision of classification by adding labels, the accuracy across all three algorithms decreases predictably. From the figures, we see that the Bayes Network performs worse than the other two classifiers. This is attributed to the simplified assumption that the features that we use are not conditionally independent. For example, the motion sensors associated with an activity are used to find the total number of motion sensor events was triggered and also the kinds of motion sensors involved in the activity.

The ROC curve is not as strong of a measure as accuracy for this experiment. That is because the accuracy is difficult to measure when the training dataset is highly skewed. To deal with these imbalanced datasets, we apply the SMOTE sampling to rebalance our datasets by increasing the number of minority class instances, thereby enabling the classifiers to learn more relevant rules for the minority class. Figure 14 depicts the new class distribution of the Kyoto and Tulum datasets after applying the SMOTE. From the figure, we see that two other minority classes have been increased greatly after balancing the datasets.

To analyze the effectiveness of the sampling technique, we evaluate the accuracy of a SVM prediction algorithm on both datasets with and without the sampling. The results are shown in Figures 15 and 16. In Figure 15, the accuracy for both datasets decreased slightly. On the contrary, Figure 16 depicts the performance as measured by the area under the ROC curve, which has been improved. After sampling, the classifiers improved the performance to classify the minority class with the loss of decreasing the accuracy. The experimental results show



(a) Accuracy (b) AUC

Figure 13: Comparison of the accuracy and AUC for the Tulum dataset .

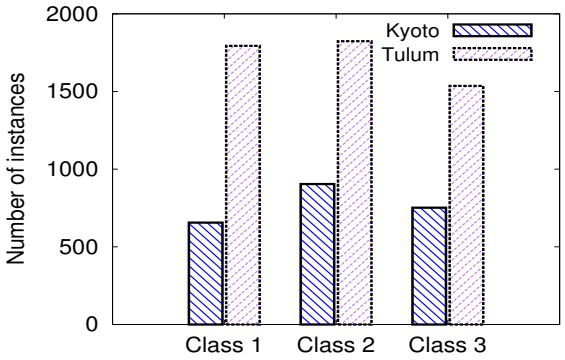
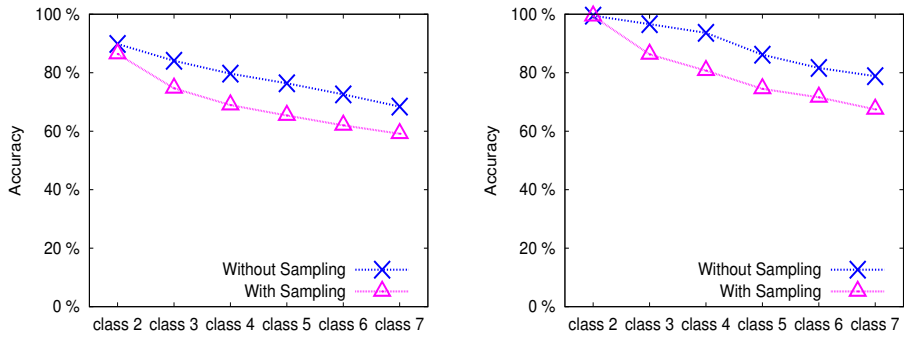


Figure 14: New distribution of instances in the three energy classes for Kyoto and Tulum.

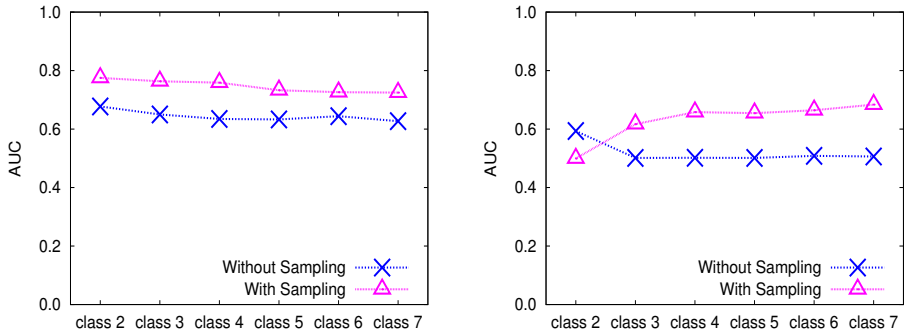
that the sampling technique is a good approach to rebalance the energy data and further improve prediction performance on minority class.

Figures 15 and 16 also compare the performance of the support vector machine for two different environments, Kyoto and Tulum. Looking at the graphs, Tulum yields a slightly improved performance over Kyoto. This is likely due to the fact that some energy-intensive devices such as room heaters were used in Kyoto but not Tulum (heat is handled from a separate building source in Tulum). These devices are not under the direct control of residents, nor are they directly impacted by activities.



(a) Kyoto (b) Tulum

Figure 15: Comparison of the accuracy with and without sampling.



(a) Kyoto (b) Tulum

Figure 16: Comparison of AUC with and without sampling.

6. Mobile Sustainability Intervention

The last component of CASAS-Sustain is behavior-based intervention to promote sustainability in everyday environments. We focus on a pervasive approach to promote sustainability behavior. Figure 17 shows our CASASviz visualizer, which is web-based and can therefore run on a computer display or a mobile device [33]. CASASviz describes an environment graphically using Scalable Vector Graphics. In one mode (the one shown in Figure 17(a)), users can look at an interface which displays sensor events that occur in the environment in real time or play back mode. Because CASASviz also displays the corresponding resource utilization (power, temperature, burner, and water, as shown in Figure 17(b)), users can get a quick view of their energy consumption and the corresponding activities

(a) Track Residents

(b) Monitor Energy

Figure 17: Mobile device-based intervention for sustainable behavior.

in the environment that impacts their energy utilization.

7. Conclusions

In this article, we consider the role of in-home behaviors upon energy usage. In particular, we analyze patterns of energy usage by monitoring activities as well as collect energy usage data from several smart environments. We analyzed the energy patterns by identifying frequent sequences of energy usage ranges and identifying outliers in the data. We further identify the role of behaviors for energy usage by using machine learning methods to map activities performed in the environment with their corresponding energy usage. Finally, we proposed a method to use this information as the basis of an intervention that allows individuals to perform their daily activities in a more energy-efficient manner. All of our algorithms were evaluated based on data collected in the CASAS smart environment testbeds.

The current state-of-the-art approaches for informing residents about their power consumption are limited. They either provide information at too low resolution through monthly totals, or at too high resolution through instantaneous measurements. These tools put the numbers and costs in context by associating energy with recognizable behaviors. It is this association that can be used to inform people about their impact on the world in a way that they can use to change their own day to day activities. The purpose of this study is to validate our hypothesis that energy usage can be analyzed and predicted based on the sensor data

that is generated by the residents in a smart home environment. The results of this work can be used to provide feedback about a resident's energy consumption as it relates to various activities. In addition, predicted electricity use can form the basis for automating activities in a manner that consumes fewer resources, including power usage. By detecting trends and anomalies, we can find some extreme energy usage values, which may indicate blackout situations, devices that were mistakenly left on, or that may lead to potential security problems in the smart environment.

Based on our experimental analysis we found that many of these techniques are useful in highlighting data collection issues and behavioral patterns that can affect energy consumption. The link between smart home sensors, machine learning algorithms, and whole-home power usage provided these insights that would otherwise likely not be caught. Additionally, our algorithms are able to show residents how much power their daily behaviors are using. This lays the groundwork for evaluating how effective this information is at influencing activity behavior over time in an attempt to reduce power consumption.

In our ongoing work, we plan to investigate methods to detect a greater range of anomalies. We also plan to install more sensitive power meters in order to capture more accurate changes in energy consumption. To extend our existing work, we will data in a greater variety of households, which will allow us to determine whether energy predictions, energy usage trends, and energy anomalies exist and generalize across multiple settings. To save cost, we will look for a minimally instrumented sensor environment to do the same evaluations of our algorithms. Finally, we will evaluate our mobile device application to determine user acceptance of the interface and quantify the short-term and long-term effects on sustainable behaviors and energy consumption.

References

[1] Multiple, Annual energy review 2010, Technical Report, U.S. Energy Information Administration, 2011.

[2] T. Ichinose, K. Shimodozono, K. Hanaki, Impact of anthropogenic heat on urban climate in tokyo, Atmospheric Environment 33 (1999) 3897–3909.

[3] Multiple, Energy Efficiency in Buildings, http://www.wbscd.org/, 2009.

[4] S. Darby, The effectiveness of feedback on energy consumption, A Review for DEFRA of the Literature on Metering, Billing and Displays (2006).

[5] J. Seryak, K. Kissock, Occupancy and behavioral affects on residential energy use, in: Proceedings of the Solar Conference, 2003, pp. 717–722.

[6] S. Szewcyzk, K. Dwan, B. Minor, B. Swedlove, D. Cook, Annotating smart environment sensor data for activity learning, Technology and Health Care 17 (2009) 161–169.

[7] S. Darby, Smart metering: what potential for householder engagement?, Building Research & Information 38 (2010) 442–457.

[8] Y. Riche, J. Dodge, R. A. Metoyer, Studying always-on electricity feedback in the home, in: Proceedings of the 28th international conference on Human factors in computing systems, 2010, pp. 1995–1998.

[9] D. J. Cook, S. K. Das, Smart environments: technologies, protocols, and applications, Wiley Series on Parallel and Distributed Computing, Wiley-Interscience, 2004.

[10] G. Hart, Nonintrusive appliance load monitoring, Proceedings of the IEEE 80 (1992) 1870 –1891.

[11] M. Berges, E. Goldman, H. Matthews, L. Soibelman, Learning systems for electric consumption of buildings, in: ASCE International Workshop on Computing in Civil Engineering, 2009.

[12] S. Gupta, M. S. Reynolds, S. N. Patel, Electrisense: single-point sensing using emi for electrical event detection and classification in the home, in: Proceedings of the 12th ACM international conference on Ubiquitous computing, 2010, pp. 139–148.

[13] Google, Google Power Meter, http://www.google.com/powermeter/about/, 2012.

[14] Microsoft, Microsoft Hohm, http://www.microsoft-hohm.com/, 2012.

[15] Opower, Opower, http://opower.com/, 2012.

[16] J. Z. Kolter, J. F. Jr., A Large-Scale study on predicting and contextualizing building energy usage, in: the Conference on Artificial Intelligence (AAAI), Special Track on Computational Sustainability and AI, 2011.

[17] G. Bellala, M. Marwah, M. Arlitt, G. Lyon, C. E. Bash, Towards an understanding of Campus-Scale power consumption, in: Proceedings of the 3rd ACM Workshop On Embedded Sensing Systems For Energy-Efficiency In Buildings, 2011, p. 6.

[18] J. E. Seem, Using intelligent data analysis to detect abnormal energy consumption in buildings, Energy and Buildings 39 (2007) 52–58.

[19] X. Li, C. Bowers, T. Schnier, Classification of energy consumption in buildings with outlier detection, IEEE Transactions on Industrial Electronics 57 (2010) 3639 –3644.

[20] D. Cook, M. Schmitter-Edgecombe, A. Crandall, C. Sanders, B. Thomas, Collecting and disseminating smart home sensor data in the CASAS project, Proceedings of the CHI Workshop on Developing Shared Home Behavior Datasets to Advance HCI and Ubiquitous Computing Research (2009).

[21] H. Liu, F. Hussain, C. L. Tan, M. Dash, Discretization: An enabling technique, Data Mining and Knowledge Discovery 6 (2002) 393–423.

[22] D. Gusfield, Algorithms on strings, trees, and sequences: computer science and computational biology, Cambridge University Press, 1997.

[23] E. M. McCreight, A space-economical suffix tree construction algorithm, Journal of the ACM 23 (1976) 262–272.

[24] J. Tukey, Exploratory data analysis, Addison-Wesley, Reading, 1977.

[25] E. Kim, S. Helal, D. Cook, Human activity recognition and pattern discovery, IEEE Pervasive Computing 9 (2010) 48–53.

[26] I. Coope, Circle fitting by linear and nonlinear least squares, Journal of Optimization Theory and Applications 76 (1993) 381–388.

[27] J. Pearl, Probabilistic reasoning in intelligent systems: networks of plausible inference, Morgan Kaufmann, 1988.

[28] B. E. Boser, I. M. Guyon, V. N. Vapnik, A training algorithm for optimal margin classifiers, in: Proceedings of the annual workshop on Computational learning theory, 1992, pp. 144–152.

[29] S. Zornetzer, An introduction to neural and electronic networks, Morgan Kaufmann, 1995.

[30] N. V. Chawla, N. Japkowicz, A. Kotcz, Special issue on learning from imbalanced data sets, ACM SIGKDD Explorations Newsletter 6 (2004) 1–6.

[31] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, Journal of Artificial Intelligence Research 16 (2002) 321–357.

[32] I. H. Witten, E. Frank, Data mining: practical machine learning tools and techniques, Morgan Kaufmann Pub, 2005.

[33] C. Chen, P. Dawadi, CASASviz: Web-based visualization of behavior patterns in smart environments, in: IEEE International Conference on Pervasive Computing and Communications, 2011, pp. 650–652.