Outlier Detection in Smart Environment Structured Power Datasets

Vikramaditya Jakkula
Department of E.E.C.S., Washington State University,
Spokane Street,
Washington State University,
Pullman, WA 99164-2752

Diane Cook
Department of E.E.C.S., Washington State University,
Spokane Street,
Washington State University,
Pullman, WA 99164-2752

Abstract—Household electricity consumption is a direct contributor to household expenses. Electricity acts as a backbone for a strong economy [1]. The rise in the energy consumption is clearly observed in this past decade, and so is the rise in the need for energy efficiency and conservation [2]. Monitoring power consumption by using various devices and instruments is on the rise; however a smart environment scenario needs more than just real-time monitoring. The need for identifying abnormal power consumption is clearly present. In this paper, we introduce our work on building novel outlier detection algorithms which uses statistical techniques to identify outliers and anomalies in power datasets collected in smart environments. We also experiment clustering techniques on the same dataset and report the results found.

Outlier Analysis, Smart Environments, Statistical Analysis, and, Data Mining.

I. INTRODUCTION

Household electricity consumption is a direct contributor to the household expenses. Electricity acts as a backbone for a strong economy [1]. We have certainly witnessed this, since the beginning of electrification in the world; household consumption has been on the rise. Recent trends and events are forcing a much slower rate in household consumption, and thus we witness the increasingly growing household consumption over the last decade.

There are various factors which contribute directly or indirectly to the increase of the consumption. The recent changes in climatic conditions and the increased usage in central electric heating and cooling systems, in the last decade, brought new levels of comfort and convenience to consumers, as well as sharp increases in consumption. In addition to electric heating and cooling, increases in the ownership and use of electric water heaters, electric kitchen appliances, refrigerators, televisions, and various electronic devices have contributed to the rise in household energy use. Home size is also a contributing factor. When it comes to homes, bigger is often considered to be better, thus increasing consumption. There are other factors which limit the growth such as the declining number of people per household, economic recession, and so forth. But these are short lived and do not cause great impact to the rising household energy consumption. As a result, much of global consumption growth in the next decades will continue to rise. Secondly, many developing countries already face a problem in meeting demand due to lack of capital to build capacity, improve efficiency, and thus the need to conserve grows drastically.

There are many means to conserve power; one primary way is to use energy preserving devices at home. Secondly, monitoring consumption can help reduce usage of unnecessary devices at home and thus save power. However, enhancing real-time monitoring with outlier detection and feedback can reduce electricity use drastically.

In this paper, we present our preliminary effort to investigate and build an outlier detection mechanism to enhance smart environment to conserve power, and make them efficient. In particular, we propose to use statistical methods to address outlier detection and report the findings. This paper concludes with reporting the findings. There are many benefits for this research and the outlier detection in power datasets is of great significance. Some general scenarios include, this framework can act as an early indicator for abnormality in the working of smart home, increased power consumption can mean increased resident activity or higher power consumption thus helping then change their way of life, and finally, help warn the resident of device failures and aid to save them from potential hazards.

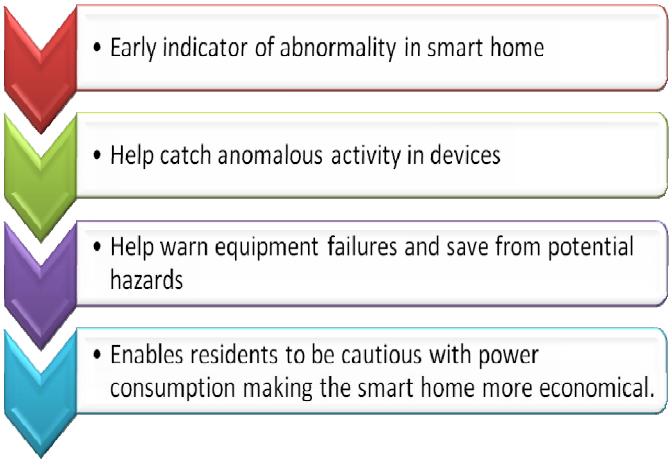


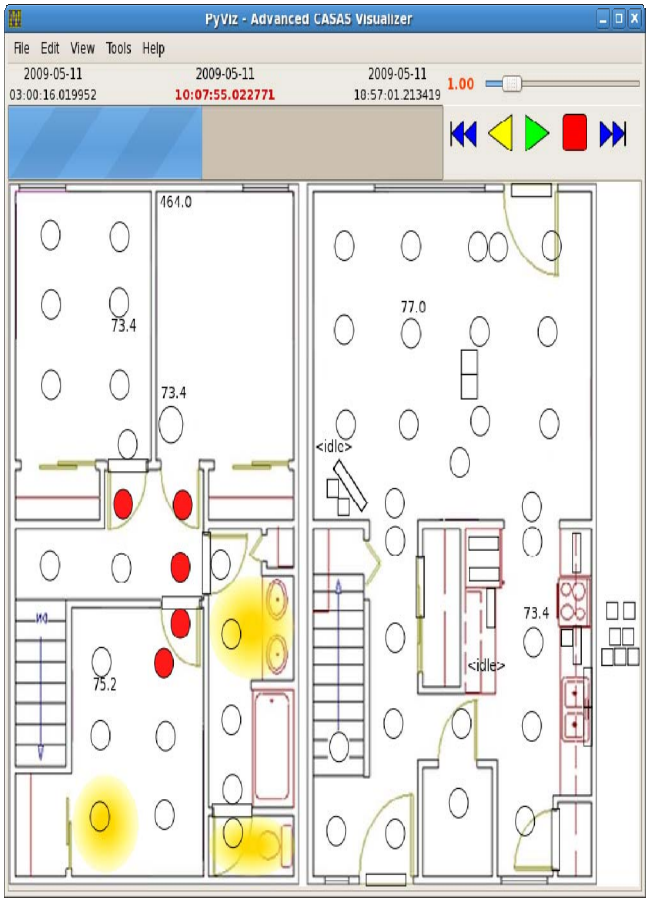
Figure 1. Significance and benefits.

II. Related Work

Energy is one of the most critical resources which encompass oil, gas, hydro-power and uranium used to generate electricity, which in turn is consumed in homes and businesses [3]. There is tremendous research being performed in smart grid related research for electricity conservation. However, there is little but promising research being done in smart home energy conservation as this is a fairly new direction of work in smart environment research due to its increasing prominence. Projects with physical test beds include the MavHome project [4], the Gator Tech Smart House [5], the iDorm [6], and the Georgia Tech Aware Home [7]. Smart home related research labs across the world have recently began investigation in this direction during the last few years. On the other hand, industrial research is rapidly progressing as well with Google's power meter and Microsoft hOhm, are definite signs of industrial interest in these areas.

III. TEST Data

Datasets are critical in order to test, compare, and enhance experimentation. For our experimentation, we use power or electricity dataset collected in the CASAS smart environment [8]. We also use synthetic data which is generated and injected with outliers via scenarios. The dataset comprises of date time and the power reading. The synthetic dataset is similar to the real dataset in terms of attributes collected.



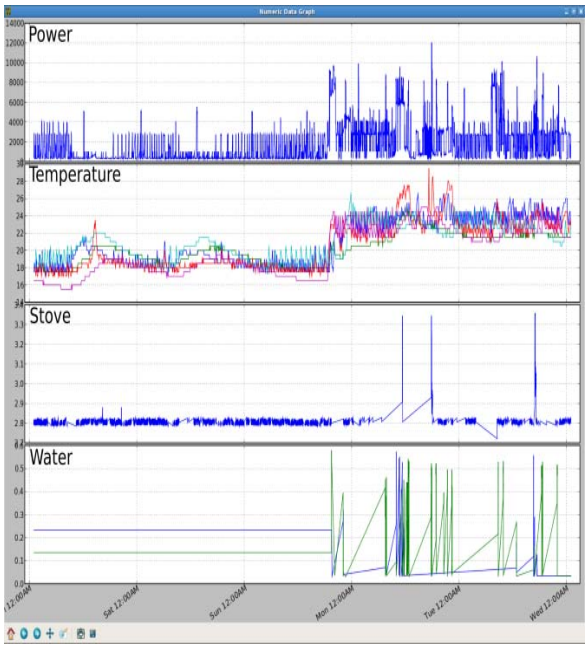


Figure. 2. Project CASAS: Physical test bed for smart environment data collection. The bottom portion of the figure displays power usage as well as readings for temperature, stove use, and water use while sensor events are generated (shown as red circles) in the testbed environment.The power data was collected by a simple power meter device.

The real dataset consists of data for a three month period. The data is collected as an hourly reading of electricity consumption in a home. The synthetic data is randomly-generated data with outliers injected by scenarios during synthetic data generation. These scenarios could be as simple as a device being turned on and forgot would increase the power consumption for a single day, which is only checked the following day, another example would be that of a running a toaster till its burns out, and, so forth. We generated synthetic data to simulate a period of twelve months with outliers included via scenarios. These scenarios are predetermined and fed into the synthetic data generation. We can vary outlier's for different experiments. Some statistics about the datasets are presented in Table 1.

TABLE I. DATASETS. WE PRESENT DETAILS ABOUT THE STATISTICS OF DATA USED IN THE EXPERIMENTS.

Dataset	Total Data points	Duration Collected/Generated	Missing Data
Real	1644	3 Month	N/A
Synthetic	8748	12 Month	N/A

Figure 3 below illustrates as how the a typical real world power reading dataset looks like and is being compared to that of a synthetic one. The plot above depicts a typical real world power reading, we can note a single outlier visually (This is seen as the plot in green color). The plot below in the figure 3. depicts that of a synthetic dataset with injected outliers (This is seen as the plot in red color).

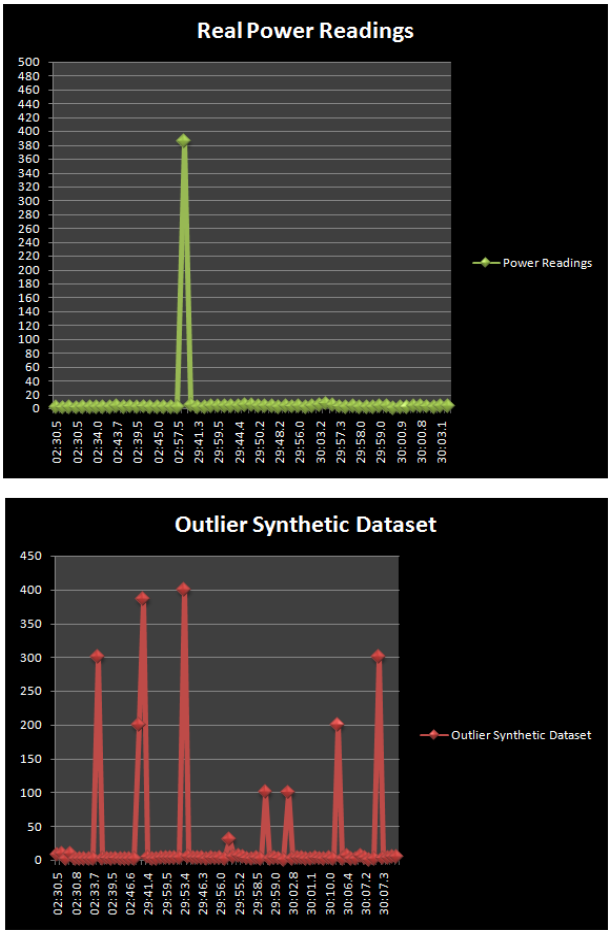


Figure. 3. Real data sets vs. Synthetic data sets. The figure highlights and compares real world data with synthetic data.

IV. EXPERIMENTATION

Outliers are observations in the data that have extreme values relative to other observations observed under the same conditions. Observations may be outliers because of a single large or small value of one variable or because of an unusual combination of values of two or more variables [9]. There are many elements which act as sources of outliers, such as data errors, implausible values, and "rare" events. Developing techniques to look for outliers and understanding how they impact data analysis are extremely important, especially in adaptive systems, such as smart environments. In the presence of outliers, any system can be distorted and would hinder performance. In smart environments, outliers hinder adapting to inhabitants or residents. Handling outliers would thus improve the overall efficiency and performance of the smart home systems. In our work we test our algorithms on energy datasets collected in smart homes and perform outlier analysis on this data.

A. Outlier Detection: Statistical Approach

In this experiment we build a statistical algorithm to observe the power datasets and filter out any outliers. Outliers are extreme data elements and the statistical approach is ideal to filter such extreme power reading, thus alerting the home resident to take appropriate action. The algorithm uses a statistical approach and identifies the outliers by identifying the extremes using standard deviation. We should note that we are using t-distribution (t-score) because we are uncertain about the results from the standard deviation from the samples (we are selecting a sample from a total population and have the moving window). In this experiment, the window size is used to pick local samples from entire population, imagine the seasonal activities in a home and the window size could be set accordingly. We also include mechanism to rank the identified outliers to measure severity of the outlier.

TABLE II. PSEUDO CODE ILLUSTRATING THE STATISTICAL APPROACH FOR OUTLIER DETECTION.

Step 1:
Define Set of outlier = {} // start with empty.

Step 2:
For i = 0 to winSize // winSize is defined by the experimenter, helps as moving window.
Compute average (A) of elements in set X
Compute Standard Deviation(S) of Elements in set X
If S>0
Find i extreme in X // we can use standard deviation based distance measures here.
Calculate standardized element: R (i) =x (I)-A/S
Calculate critical value of i: alpha
alpha (i) = (n- no of outliers) *t-score of limited sample/ ((n-outliers-1) * t-score of limited sample) ^1/2
If R (i) >alpha (i) then outlier
Remove i from set; i is added to the outlier set
We continue till end is reached.

Step 3: Ranking (set of outliers);

We measure ranking by the severity of the outlier and extreme severe ones can be considered as anomalies. In all the outliers identified simply find standard deviation and standard deviation +1, standard deviation +2, standard deviation +3 can be used for ranking.

B. Outlier Detection: Clustering Approach

Outlier detection techniques were once very naïve, but over the course of time have evolved into more systematic procedures. In the second experiment we cluster the datasets to identify potential outliers. Generally outlier detection involves distance based, distribution based or density based. The clustering effort helps us compare these two models. For clustering we use K-Nearest Neighbor (KNN) algorithms where Discrete Time Warping (DTW) is used for distance measure. KNN will be successful, because the point-to-point distance measure will distinguish the anomalous data from normal data. Additionally, DTW will align the power values and hence improves the results. The DTW algorithm has earned its popularity by being extremely efficient as a time-series similarity measure, given our data is collected against time. Thus it is a better choice than general distance measures such as Euclidean, Manhattan, and Chebyshev measures.

TABLE III. PSEUDO CODE FOR KNN [10] AND DTW FOR DISTANCE MEASURE [11]

```
For each element X in the data do
If X has unknown distance or missing data
 X is abnormal
Else then
 For each element in data
  Calculate distance
 Find the biggest score of X in distance measure
 Calculate the average distance for k nearest neighbors
  If average distance is greater than threshold then
 X is normal
 Else then
 X is abnormal

Declare DTW [0.n, 0.m]
Declare i, j, cost
   For i: = 1 to m
      DTW [0, i]:= infinity
   For i: = 1 to n
      DTW [i, 0]:= infinity
   DTW [0, 0]:= 0
   For i: = 1 to n
      For j: = 1 to m
        Cost: = d(s[i], t[j])
 DTW [i, j]:= cost + minimum (DTW [i-1, j], // insertion
 DTW [i, j-1],    // deletion
 DTW [i-1, j-1])    // match
 Return DTW [n, m]
```

C. *Outlier Detection: Experimentation Process*

The first experiment compares outlier analysis on real and synthetic datasets using the above two suggested approaches and reports the results. The second experiment consists of varying the outliers, and seeing if the techniques identify them correctly and, thus measure how accuracy varies. We have varied the outliers by increasing the number of outliers present in the synthetic dataset. This experiment uses synthetic datasets only.

V. DISCUSSION AND OBSERVATION

In the first experimentation process we run both techniques on the real and synthetic datasets and present the results obtained. These results are displayed in Table 2. The clustering approach outperforms the statistical approach with outlier identification

TABLE IV. EXPERIMENTATION RESULTS OVERVIEW.

Dataset	Total Data points	Actual Outliers	Outliers Identified via statistical approach	Outliers Identified via clustering
Real	1644	N/A	86	2
Synthetic	8748	16	104	16

The second experimentation process consists of a statistical approach and a clustering based approach on synthetic datasets with a varying number of outliers. The statistical approach looked promising on smaller outlier sets, and as the number of outliers increased in the dataset the performance degrades. The clustering based approach has a steady performance while outlier numbers are varied. The current experiment is performed on limited datasets, but in on future work we propose to perform the same analysis on larger real datasets. Table 3 displays the observations for this experiment.

TABLE IV. DETAILED EXPERIMENTATION RESULTS.

Experiment Process	# Outliers in Synthetic Data	Outliers Identified	Correctly Identified	Incorrectly Identified	Accuracy (%)
Outliers Identified via statistical approach	2	2	1	1	50%
	5	3	3	0	100%
	50	39	12	17	30.70%
	80	52	28	24	35.00%
	100	69	34	35	34.00%
Outliers Identified via clustering	2	2	2	0	100%
	5	5	5	0	100%
	50	47	47	0	94.00%
	80	82	69	13	92.50%
	100	110	86	24	92%

VI. FUTURE WORK

Comprehensive energy outlier analysis and fault detection system is an interesting area to research and develop. The next step is to continue the work with larger datasets and work to improve the accuracy. We propose to strengthen the experimentation with a layered approach, where we have an ensemble of classifiers help to identify outliers.

In this layered approach we create an outlier holding set. Layer 1 will use a statistical approach to identify outliers and add them to the holding set. Layer 2 will use events as well as activity-based abnormality detection and will add the corresponding detected outliers to the holding set. Finally, Layer 3 will identify user/resident-indicated outliers or abnormal activity. Our automated approach will clean the outlier set of duplicates and will rank outliers based on significance and criticality.

Some of our near future work will include correlation analysis to identify smart home attributes which have a direct impact in power electricity consumption. For example, whether the numbers of devices in the house have any correlation with the daily power consumption? Some future work will involve providing forecasts of household electricity consumption using neural networks, markov models, and support vector machine techniques. Building predictive models for household electricity energy consumption will also be evaluated. The fusion of external factors such as changing weather against time will also be investigated. There are ideas to include outlier detection systems as a component in the consumption prediction technique, thus making it act as a reinforcement

mechanism for consumption prediction leveraging outlier detection.

Privacy is another important concern today. Privacy adds a whole new dimension to power monitoring systems today. Privacy and personalization are two areas of huge interest for overall power monitoring systems research. On the other hand such devices should have strong security layer to prevent any collected or computed data from being made available public so as to protect the user's privacy. We envision our system to grow beyond an outlier detection system into a true power monitoring system empowered with smart suggestion and goal setting features.

VII. CONCLUSIONS

Real-time smart power monitoring system based on outlier detection means real savings for the household resident. Despite the difficulties, exploring why outliers exist can provide many clues to the development of better models, especially for adaptive systems. In fact, well designed and good performing systems enable smart home residents to follow and usher an era of trust worthy computing. In this paper, we present our work as an introduction to the problems of outlier analysis in power datasets from smart environments, and also present their detection, and approaches to data analysis through experimentation process. The experiment results present stable results, and justify our preliminary effort in building outlier detection framework for energy datasets. We observe that the clustering based approach appears to be more successful than a mere statistical analysis on the datasets. We propose to further extend this work further, and explore layered approaches for an outlier detection framework.

REFERENCES

[1] Energy Star web site, http://www.energystar.gov/index.cfm?c=cfls.pr_cfls

[2] Energy Information Administration, Annual Energy Outlook 2008, page 61.

[3] Li, X., Bowers, C.P. & Schnier, T. Classification of Energy Consumption in Buildings with Outlier Detection. IEEE Transactions on Industrial Electronics (2009)

[4] G. M. Youngblood and D. Cook. Data mining for hierarchical model creation. IEEE Transactions on Systems, Man, and Cybernetics, Part C, 37(2007), 1-12.

[5] A. Helal, W. Mann, H. El-Zabadani, J. King, Y. Kaddoura, and E. Jansen. The Gator Tech smart house: A programmable pervasive space. IEEE Computer, 38(2005), 50-60.

[6] F. Doctor, H. Hagras, and V. Callaghan. A fuzzy embedded agent-based approach for realizing ambient intelligence in intelligent inhabited environments. IEEE Transactions on Systems, Man, and Cybernetics, Part A, 35(2005), 55-56.

[7] G. Abowd and E. Mynatt. Designing for the human experience in smart environments. In D. Cook and S. Das, editors, Smart Environments: Technology, Protocols, and Applications, Wiley, 2004, pp. 153-174.

[8] D. Cook, M. Schmitter-Edgecombe, A. Crandall, C. Sanders, and 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.

[9] Mason, Robert L.; Gunst, Richard F.; Hess, James L. Statistical Design and Analysis of Experiments - With Applications to Engineering and Science (2nd Edition). John Wiley & Sons.

[10] Wikipedia. k-nearest neighbor algorithm. http://en.wikipedia.org/wiki/K-nearest_neighbor_algorithm. Retrieved November 10, 2009.

[11] Wikipedia.DTW. http://en.wikipedia.org/wiki/Dynamic_time_warping. Retrieved November 10, 2009.