Association Rule Mining of Household Electrical Energy Usage

Devika Bhagwandin
The University of the West Indies
St. Augustine, Trinidad and Tobago
devikabhagwandin@hotmail.com

Vrijesh Tripathi
The University of the West Indies
St. Augustine, Trinidad and Tobago
Vrijesh.Tripathi@sta.uwi.edu

Patrick Hosein
The University of the West Indies
St. Augustine, Trinidad and Tobago
Patrick.Hosein@sta.uwi.edu

ABSTRACT

In a Smart Grid network, household energy consumption is monitored by the use of residential smart meters. However, these do not provide data for individual appliance consumption. In this paper we consider publicly available dis-aggregated electrical energy data obtained from six homes. Using the apriori algorithm we determine association rules between appliances so as to understand which appliances are used in conjunction with others. This information can then be used to determine better ways to control the on/off activity of these devices. Our objective was to determine if such associations can in fact be determined. We conclude that this is indeed the case as demonstrated by our results. In general, houses with a large number of appliances generated more association rules while the accuracy of classifying rules reduced with an increase in the number of appliances.

KEYWORDS

Association Mining, Apriori algorithm, Data Mining, Smart Grid

ACM Reference format:
Devika Bhagwandin, Vrijesh Tripathi, and Patrick Hosein. 2017. Association Rule Mining of Household Electrical Energy Usage. In *Proceedings of DMCIT '17, Phuket , Thailand, May 25-27, 2017,* 5 pages.
DOI: http://dx.doi.org/10.1145/3089871.3089888

1 INTRODUCTION

Electricity is a form of energy used on a daily basis from small scale applications at home, to large scale applications in factories. It is produced in power stations from both renewable and non-renewable primary sources of energy including fossil fuels, wind, water, solar and nuclear energy [8]. This electricity is then distributed to homes and workplaces for consumption. Although electricity is essential, studies have shown that commercial, industrial and residential energy consumption may result in increased emission of greenhouse gases. Controlling greenhouse gases in the atmosphere is important to reduce the impact of climate change [20]. While significant research has been done on making energy production more efficient, more work needs to be done on improving the efficiency of its distribution and usage.

The United States is the world's second largest energy consumer, China being the first. In 2014, 18% of the world's primary energy was consumed by the United States amounting to 98 quadrillion British thermal units (BTU) [23]. In 2015, 20558 trillion BTUs were consumed by residential end users [22]. This sector has huge potential for initiating strategic use of electricity as users can easily monitor energy usage at home. Being energy efficient can benefit individuals and countries in several ways. It can lead to dynamic pricing of domestic electricity for customers deciding to use appliances when it is cheaper to use it.

Smart meters are traditionally used to monitor electricity consumption. These meters however, only provide the whole home power consumption. Such data is not really useful to the end user if an individual were to try to become energy efficient. The consumer needs knowledge of which appliances contribute the most to the electricity bill. Once users know where energy is being wasted, steps can be taken to reduce unnecessary electricity consumption. This is where energy dis-aggregation is useful. Energy dis-aggregation is the process of extracting end-use and appliance-level data from an aggregate or whole building usage [3]. A common method for energy dis-aggregation is Non-Intrusive Load Monitoring (NILM). NILM requires one single meter for the house being monitored and detects changes in current and voltage deducing the appliances used in that house and their respective energy consumption [26].

1.1 Related Work

Past studies including [10, 11, 14] and [15] focus on the collection and analysis of residential dis-aggregated energy data sets. The methods used to analyze the respective data sets include Discriminate dis-aggregation Sparse Coding (DDSC), Factorial Hidden Markov Models (FHMM), Marked Hawkes process and time series. These studies develop energy dis-aggregation algorithms to reduce training requirements and improve its robustness and accuracy. In another study [18], rule mining is used to compare the strength of time-based associations to the associations between devices by means of a JMeasure metric.

1.2 Contributions

The main difference between this report and past studies is our use of the apriori algorithm in association rule mining to examine the occurrence of appliances: both concomitant and single. Such analysis provides the necessary information to improve on residential energy conservation. Consumers will now be aware of the appliances that are used together and the appliances which are on but not necessarily in use (idling). Our main contribution is the demonstration that Association Mining can in fact be used to find association rules with sufficiently high confidence to be useful for managing home appliances and devices.

DMCIT '17, Phuket , Thailand

DOI: http://dx.doi.org/10.1145/3089871.3089888

2 DATA ANALYSIS

2.1 The REDD Data Set

The publicly available Reference Energy dis-aggregation Dataset (REDD) was obtained from [12]. This data was collected to assist researchers in the field of data mining and machine learning. The data was collected from six houses across the greater Boston area in the United States and consists of low frequency, high frequency and raw high frequency measurements. The measurements were collected by observing and monitoring the homes over a period of several months. The exact period of time was not made available due to reasons of confidentiality. The data is therefore supervised information. Hardware and Software systems were setup across each home and these are further explained in [11]. The total home energy usage was recorded at a frequency of 15 kHz whereas the appliance loads were sampled at a rate of 1 Hz.

In this study only the low frequency (appliance) measurements were analyzed. The samples were time stamped with UTC time. The first two sample channels of each house were the mains, set at a frequency of once per second, and the remaining channels were appliance circuits set at a frequency of once every three seconds. Houses one to six contained 20, 11, 22, 20, 26 and 17 channels respectively, including the mains.

Appliances that were not on at any time during the period under study were removed. The resulting numbers of channels were 18, 11, 22, 19, 21 and 16 for houses one through six respectively. This study focuses on the appliances so the two mains for each house were not included in the analysis. Table 1 provides a list and number of appliances for each house.

Table 1: List of Appliances for Each House

House	Appliances	Number
1	oven1, oven2, refrigerator, dishwasher, kitchen outlet1, kitchen outlet2, lighting1, washer-dryer, microwave, bathroom gfi, electric heat, stove, kitchen outlet3, kitchen outlet4, lighting2, lighting3	16
2	kitchen outlet1, lighting, stove, microwave, washer-dryer, kitchen outlet2, refrigerator, dishwasher, disposal	9
3	unknown outlet1, unknown outlet2, lighting1, electronics, refrigerator, disposal, dishwasher, furnace, lighting2, unknown outlet3, washer-dryer1, washer-dryer2, lighting3, microwave, lighting4, smoke alarm, lighting5, bathroom gfi, kitchen outelt1, kitchen outlet2	20
4	lighting1, furnace, kitchen outlet1, unknown outlet, washer-dryer, stove, air conditioning1, air conditioning2, miscellaneous, smoke alarm, lighting2, kitchen outlet2, dishwasher, bathroom gfi1, bathroom gfi1, lighting3, lighting4	17
5	microwave, unknown outlet1, furnace, unknown outlet2, washer-dryer1, washer-dryer2, subpanel1, subpanel2, electric heat1, electric heat2, lighting1, unknown outlet3, bathroom gfi, refrigerator, lighting2, dishwasher, electronics, lighting3, kitchen outlet	19
6	kitchen outlet1, washer-dryer, stove, bathroom gfi, refrigerator, dishwasher, unknown outlet1, unknown outlet2, electric heat, kitchen outlet2, lighting, air conditioning1, air conditioning2, air conditioning3	14

2.2 Association Rule Mining

Association Rule Mining or Association Rule Learning is a technique in data mining used to discover relationships or associations among variables [21]. Note that we only consider appliance associations within each house but do not consider associations across houses (e.g., whether a dishwasher in house 1 is associated with a dishwasher in house 2).

Consider a single house and let T denote the set of active devices at each sampling time so that $t \in T$ represents the set of active appliances at a given sampling point. Let X denote some subset of appliances and let Y denote another subset. These two sets (called itemsets) are said to be associated if X and Y have one or more common items. An association rule denoted by $X \rightarrow Y$ expresses an if/then relationship between sets X and Y. The *Support* of X with respect to T is defined as the fraction of all instances in which the items in X are active:

$$\text{supp}\{X\} \equiv \frac{|t \in T; X \subseteq t|}{|T|} \tag{1}$$

The *Confidence* value of an association rule $\{X \rightarrow Y\}$ represents how often the rule has been found to be true:

$$\text{conf}\{X \rightarrow Y\} \equiv \frac{\text{supp}(X \cup Y)}{\text{supp}(X)} \tag{2}$$

Finally the *Lift* of an association rule is the ratio of the observed support to that expected if the sets were independent:

$$\text{lift}\{X \rightarrow Y\} \equiv \frac{\text{supp}(X \cup Y)}{\text{supp}(X) \times \text{supp}(Y)} \tag{3}$$

If X and Y occur independently then the Lift is 1 otherwise it increases as the correlation increases [4, 5, 13].

The Apriori algorithm (described in [1, 9, 17]) is used for our association analysis. This algorithm determines those itemsets of appliances that occur frequently. Given these frequent itemsets we then determine association rules that have high confidence values (i.e. rules that occur frequently). If association rule $X \rightarrow Y$ occurs frequently then we can say (with some degree of confidence) that whenever items in X are active then items in Y are also active. This then tells us the relationship between active devices in a home.

Association Rule Mining was conducted using the Statistical Software **R**. The **R** packages "arules" and "arulesViz" were used to analyze associations between all appliances separately for each of the six houses [6, 24]. Note that one could also analyze associations across houses and this may be done in the future. From the dataset we can determine, for each time sample, the set of devices which are on since their power consumption would be positive. Given this information we can form the set T, execute the Apriori algorithm to determine the frequent itemsets and finally generate the association rules. The generated rules are then sorted, inspected and its redundant rules removed. These resulting rules are then displayed graphically.

2.3 Validation

Sensitivity is defined as the probability that a positive (or significant) rule is classified correctly (i.e., positive correctly classified divided by total positive rules), whereas *specificity* is the probability that a negative (or insignificant) rule is classified incorrectly (i.e., negative incorrectly classified divided by total negative rules). *Accuracy* is the probability that the algorithm classifies both significant and insignificant rules correctly (i.e. sum of positive and negative correct classifications divided by total rules). A Receiver Operating Characteristic (ROC) curve is a plot of Sensitivity versus (1 - Specificity). It is used to evaluate the performance of the apriori algorithm. The area under the curve (AUC) is used to determine how well an algorithm classifies the rules. ROC curves with high AUC values imply that the algorithm is good at classifying the association rules [19, 25].

To validate the results, the rules generated from the apriori algorithm were classified using two parameters, support and confidence. Contingency tables for each of the houses were then constructed. The sensitivity, specificity and accuracy values for each house and for both classifications were calculated and tabulated. Finally, AUC values were determined by plotting ROC curves for each house using the calculated sensitivity and specificity values.

3 RESULTS

In this section the experimental results together with explanations are presented. With the given data we can compute energy consumption for the various appliances (energy is the product of power and the time during which the appliance is on). The appliances which consumed the most energy in houses 1 to 6 are refrigerator (house 1), refrigerator (house 2), electronics, furnace, lighting-3 and lighting respectively. It was found that house 6 consumed the most power and house 2 consumed the least. Also house 6 consumed the most energy but house 4 consumed the least. House 4 consumed the least energy but not the least power as the average on time of its appliances was smallest.

We first use scatter plots to demonstrate the various metrics for each of the generated rules. Due to space limitations we only provide this plot for House 3 which was a typical house. This plot, provided in Figure 1, contains the confidence value of the rule on the vertical axis, its support on the horizontal axis and the shade of the diamond indicates the lift (with lighter colored rules having higher lift values). It is seen that generally, across all the houses, the association rules satisfy minimum support and minimum confidence. Rules which also have both high support and high confidence are said to be strong [7, 24]. Houses 1, 2, 3, 4, 5 and 6 generated 28, 15, 168, 47, 590 and 39 significant rules respectively.

Next we provide a graph-based visualization of the association rules. Again, we only provide this for House 3 which has 168 rules. In this graph, shown in Figure 2, each rule is represented by a circle with the size of the circle and its color denoting the support and lift respectively of the rule. Larger circles denote larger support values while darker circles denote larger lift values. Therefore, the rule with smallest support and largest lift would be identified as the smallest and darkest colored circle. Notice that electronics is apart from the other rules, this suggests that it was used by itself. It is also seen that this rule has high support and low lift.

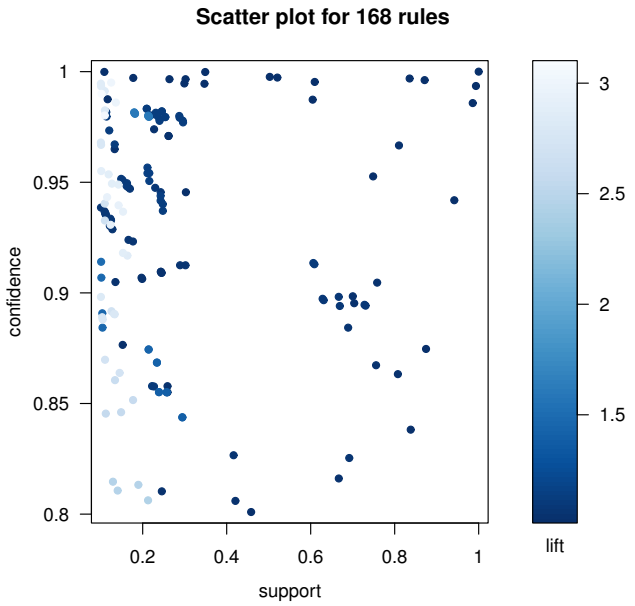

Figure 1: Significant Association Rules for House 3

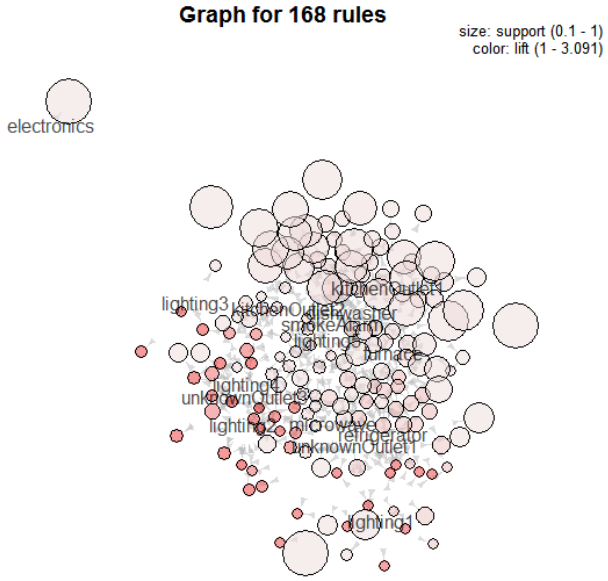

Figure 2: Graph-based Visualization of Rules for House 3

The largest association for each house was determined from the **R** output and tabulated in Table 2. It is clear that house 5, with 10 appliances working together, has the largest combination of associations among all the houses. House 3 has the second largest combination of associations, with 6 appliances working together.

Table 2: Largest Association for Each House

House	Association
1	refrigerator, bathroom gfi, kitchen outlet3 → lighting 1
2	stove, kitchen outlet 2, refrigerator → washer, dryer
3	dishwasher, furnace, microwave, smoke alarm, kitchen outlet1 → lighting 5
4	unknown outlet, stove, miscellaneous, bathroom gfi1 → bathroom gfi 2
5	microwave, unknown outlet1, subpanel 1, lighting1, bathroom gfi, refrigerator, electronics, lighting3, kitchen outlet → furnace
6	kitchen outlet1, refrigerator, kitchen outlet2 → stove

Table 3: Sensitivity, Specificity and Accuracy for Support

House	Sensitivity	Specificity	Accuracy
1	0.2857	0.9822	0.9727
2	0.3333	0.9410	0.9125
3	0.0952	0.9905	0.9676
4	0.1702	0.9285	0.9699
5	0.2322	0.7672	0.7586
6	0.2051	0.9933	0.9892

Table 4: Sensitivity, Specificity and Accuracy for Confidence

House	Sensitivity	Specificity	Accuracy
1	0.6429	0.9718	0.9673
2	0.7333	0.9016	0.8938
3	0.4107	0.9654	0.9512
4	0.7234	0.9641	0.9603
5	0.2763	0.7224	0.7513
6	0.6154	0.9860	0.9841

Note that some of these associations may seem obvious but in more complicated scenarios this may not be the case. Furthermore, association rules may change as the seasons change and hence having an automated method for determining these associations can be useful. To summarize our findings, house 5 contains 19 appliances and generates 590 significant rules. House 3 contains 20 appliances and generates 168 significant rules. The other houses, which contain fewer appliances, generate a smaller number of association rules. The combination of associations is largest for house 5 followed by house 3. As one would expect, houses with a larger number of appliances generate more association rules and have a larger combination of associations.

3.1 Validation

The apriori algorithm was chosen since it is fast. It has also outperformed other algorithms such as AIS, SETM and AprioriTid as explained in [2]. Table 3 provides the values for sensitivity, specificity and accuracy for rule Support while Table 4 provides them for rule Confidence. Figure 3 shows the ROC curves for house 3 and Table 5 shows the area under the ROC curves for each house.

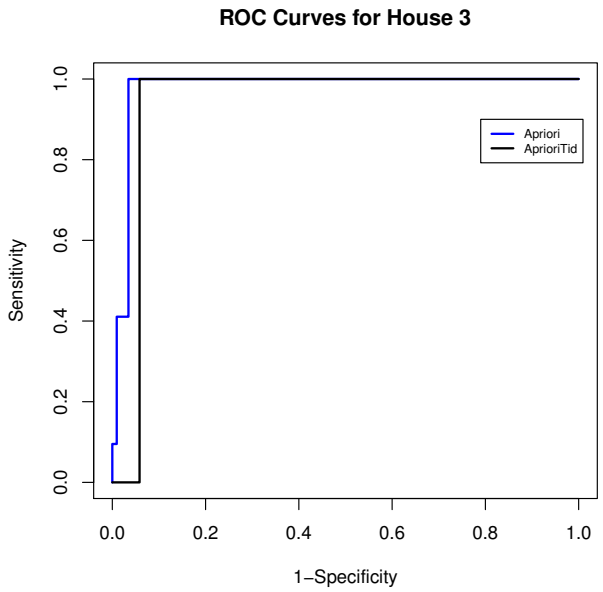

Figure 3: ROC Curves for Rule Mining Algorithms

Table 5: AUC for Rule Mining Algorithms

House	Apriori	AprioriTid
1	0.8057	0.4978
2	0.8122	0.4836
3	0.6877	0.4958
4	0.8209	0.4919
5	0.5000	0.4996
6	0.8001	0.4979

In classification by support and confidence, the sensitivity values were generally low and the specificity values were high. A high specificity measure means that a large number of false negative rules were generated. It is beneficial to have a large number of false negatives in mining frequent itemsets as it greatly reduces the quantity of significant rules generated by the algorithm thereby reducing memory consumption. The number of false negatives however, should be controlled as this could affect the sensitivity of the algorithm. The low sensitivity values suggest that a small fraction of significant rules were classified as truly positive. This may result from the large number of rules generated by the data set. To address this we suggest the use of the AprioriHybrid algorithm or FP-growth as an alternative approach in association rule mining [16] since they are better suited for large data sets.

The accuracy for houses 1, 2, 3, 4 and 6 are high. This suggests that the algorithm has a high probability of predicting rules correctly. Also, the algorithm is excellent at classifying rules for houses 1, 2, 4 and 6 since their respective AUC values are greater than 0.8. Classification of rules in houses 3 and 5 are acceptable but the areas under their ROC are not as high (between 0.5 and 0.7). These two houses have a large number of appliances. To verify that the Apriori

algorithm outperforms other algorithms, we applied the AprioriTid algorithm to the data and found its respective AUC values by plotting ROC curves. Table 5 and Figure 3 show that this is indeed the case as the AprioriTid classified rules poorly with AUC values less than 0.5 and its ROC curve is lower than that of the Apriori.

This study was conducted without knowledge of whether any of the houses contained energy efficient gadgets and without knowledge of the time period over which the data was collected. If this period was known, then conclusions would be better with regard to energy consumption in a specified season. Also, if the hour of the day the appliances were used in each house were known, it would then be possible to graph the usage of power throughout the day and determine the peak hours that power is used on a daily basis. Such data is useful as the user will be aware of his/her behavioral patterns thereby resulting in a change of their regular habits.

4 CONCLUSIONS AND FUTURE WORK

The objective of this paper was to determine if Association Mining can assist in energy savings in the home. However, because the number of houses for which samples were made available was small, we cannot make general statements. What we have demonstrated is that this technique can be used to identify which appliances are used together. One can then use this information to make changes that may help to conserve energy. For example, if we find that typically one turns on two or more lights whenever they are in a room then it is very likely that they often forget to switch off one of these lights when leaving the room. Hence having a common switch for all of the concerned lights will help to alleviate this problem leading to reduced energy consumption. One can come up with various other scenarios like this and we believe that the approach provided in this paper will assist with such solutions.

Furthermore we believe that the approach here can be used in an adaptive system. For example, in a smart home in which all lights and appliances can be centrally controlled one can do the following. Suppose we determine that the association rule $X \rightarrow Y$ has high confidence then once appliances in X are turned on, the appliances in Y can be automatically turned on. Again if we consider the case of a room with several lights and there is a close association between the activation of the first light and the other lights in the room then with a centrally controlled system the other lights can be automatically turned on when the first one is turned on. Such a system will continuously learn over time and adapt as behaviors or the environment changes. Of course, this will require that the system occasionally forget an association rule in order to determine if the rule is still valid (e.g, the person who prefers all lights on no longer lives there and the new occupant prefers one light).

One thing we did not do is determine association rules across houses. We believe that such an analysis may also lead to some interesting insights. For example, suppose we find that several people use their washing machines at the same time then this would increase the load peak and generally high peaks are undesirable in a power grid since it may require activation of additional power sources. Hence knowledge of this washing machine association rule can help in approaches such as dynamic (load based) pricing so that people avoid using their washing machine during this peak hour. One can also include additional non-energy information when determining rules. For example, given weather information, one can determine association rules with factors such as temperature. These will be addressed in future work.

REFERENCES

[1] Sakshi Aggarwal and Ritu Sindhu. 2015. *An approach to improve the efficiency of apriori algorithm.* Technical Report. PeerJ PrePrints.

[2] Rakesh Agrawal, Ramakrishnan Srikant, and others. 1994. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, Vol. 1215. 487–499.

[3] K. Carrie Armel, Abhay Gupta, Gireesh Shrimali, and Adrian Albert. 2013. Is disaggregation the holy grail of energy efficiency? The case of electricity. *Energy Policy* 52 (2013), 213 – 234. DOI: http://dx.doi.org/10.1016/j.enpol.2012.08.062 Special Section: Transition Pathways to a Low Carbon Economy.

[4] Kirk Borne. 2014. Association Rule Mining - Not Your Typical Data Science Algorithm. (2014). https://www.mapr.com/blog/association-rule-mining-not-your-typical-data-science-algorithm

[5] Michael Hahsler. 2015. A Probabilistic Comparison of Commonly Used Interest Measures for Association Rules. (2015). DOI: http://dx.doi.org/research/association_rules/measures.html

[6] Michael Hahsler and Sudheer Chelluboina. 2011. Visualizing association rules: Introduction to the R-extension package arulesViz. *R project module* (2011), 223–238.

[7] Michael Hahsler, Bettina Grün, and Kurt Hornik. 2007. Introduction to arules–mining association rules and frequent item sets. *SIGKDD Explor* 2, 4 (2007).

[8] Pennsylvania Historical and Museum Commission. 2015. Energy Resources: Introduction to Major Energy Sources. (2015). http://www.phmc.state.pa.us/portal/communities/energy/

[9] Heydar Jafarzadeh and Mehdi Sadeghzadeh. 2014. Improved Apriori Algorithm Using Fuzzy Logic. *International Journal of Advanced Research in Computer Science and Software Engineering* 4, 6 (2014), 439–447.

[10] J Zico Kolter, Siddharth Batra, and Andrew Y Ng. 2010. Energy disaggregation via discriminative sparse coding. In *Advances in Neural Information Processing Systems*. 1153–1161.

[11] J Zico Kolter and Matthew J Johnson. 2011. REDD: A public data set for energy disaggregation research. In *Workshop on Data Mining Applications in Sustainability (SIGKDD), San Diego, CA*, Vol. 25. 59–62.

[12] J. Zico Kolter and Matthew J. Johnson. 2011. REDD: The Reference Energy Disaggregation Data Set. (2011). http://redd.csail.mit.edu/

[13] K. Lai and Narciso Cerpa. 2001. Support vs Confidence in Association Rule Algorithms. In *OPTIMA* (October 10-12).

[14] Liangda Li and Hongyuan Zha. 2015. Energy Usage Behavior Modeling in Energy Disaggregation via Marked Hawkes Process.. In *AAAI*. 672–678.

[15] Om P Patri, Anand V Panangadan, Charalampos Chelmis, and Viktor K Prasanna. 2014. Extracting discriminative features for event-based electricity disaggregation. In *Technologies for Sustainability (SusTech), 2014 IEEE Conference on*. IEEE, 232–238.

[16] P Prithiviraj and R Porkodi. 2015. A comparative analysis of association rule mining algorithms in data mining: a study. *Open J. Comput. Sci. Eng. Surv* 3, 1 (2015), 98–119.

[17] Anand Rajaraman and Jeffrey David Ullman. 2011. *Mining of Massive Datasets*. Cambridge University Press, New York, NY, USA.

[18] Sami Rollins and Nilanjan Banerjee. 2014. Using rule mining to understand appliance energy consumption patterns. In *Pervasive Computing and Communications (PerCom), 2014 IEEE International Conference on*. IEEE, 29–37.

[19] Tobias Sing, Oliver Sander, Niko Beerenwinkel, Thomas Lengauer, Tobias Sing, and Oliver Sander. 2009. Visualizing the performance of scoring classifiers. *Package ROCR Version 1.0* 4 (2009).

[20] City Green Solutions. 2016. Benefits of Energy Efficiency. (2016). https://www.citygreen.ca/benefits-energy-efficiency

[21] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. 2006. *Introduction to Data Mining*. Pearson Education.

[22] U.S Energy Information Administration. 2017. Energy Consumption by Sector. (2017). https://www.eia.gov/totalenergy/data/monthly/pdf/sec2_3.pdf

[23] U.S Energy Information Administration. 2017. FAQ: What is the United States' share of world energy consumption? (2017). https://www.eia.gov/tools/faqs/faq.php?id=87&t=1

[24] Wesley. 2012. Association Rule Learning and the Apriori Algorithm. (2012). https://www.r-bloggers.com/association-rule-learning-and-the-apriori-algorithm/

[25] Wen Zhu, Nancy Zeng, Ning Wang, and others. 2010. Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS® implementations. *NESUG proceedings: health care and life sciences, Baltimore, Maryland* (2010), 1–9.

[26] Ahmed Zoha, Alexander Gluhak, Muhammad Ali Imran, and Sutharshan Rajasegarar. 2012. Non-intrusive load monitoring approaches for disaggregated energy sensing: A survey. *Sensors* 12, 12 (2012), 16838–16866.