Smart Meters Big Data : Behavioral Analytics via Incremental Data Mining and Visualization

Shailendra Singh

A thesis submitted to the Faculty of Graduate and Postdoctoral Studies

in partial fulfillment of the requirements for the degree of

MASTER OF APPLIED SCIENCE

in Electrical and Computer Engineering

School of Electrical Engineering and Computer Science

Faculty of Engineering

University of Ottawa

Ottawa, Canada

Acknowledgements

I gratefully acknowledge the unwavering support, and constant guidance of my supervisor, Professor Dr. Shervin Shirmohammadi, and I am deeply indebted for his contribution to direction and richness of research, thoughtful encouragement, careful supervision, and numerous insightful conversations during the development of the ideas for this thesis and complete course of my research. Without his inspiration I would never have been able to complete my research work.

I am also grateful to Dr. Abdulsalam Yassine for his wisdom, friendship, understanding, incisive discussions, and relentless guidance on technical and non-technical topics throughout my research. My warmest thanks go to lab manager Mr. Basim Hafidh for his kind assistance throughout my time at lab.

For financial support, I would like to express my heartfelt gratitude to Dr. Shervin Shirmohammadi and School of Electrical Engineering and Computer Science at the University of Ottawa.

Finally, I owe more than thanks to my family which includes my parents, my wife and specially my son Pratyush Singh, who has been an ever growing inspiration and motivation for me.

Dedicated to ....

my sweet son Pratyush Singh,

my wonderful parents Mrs. Viond Bala Singh and Mr. Brahm Singh, and

my magnificent wife Sushma Singh ....

Contents

List of Figures

List of Tables

List of Abbreviations

ADL	**A**ctivities of **D**aily **L**iving
AMI	**A**dvanced **M**etering **I**nfrastructure
AMPds2	**A**lmanac of **M**inutely **P**ower **d**atas**e**t Version 2
ANN	**A**rtificial **N**eural **N**etwork
AoI	**A**ppliance **o**f **I**terest
BCD	**B**ayesian **C**lustering by **D**ynamics
BIC	**B**ayesian **I**nformation **C**riterion
CBAF	**C**Cluster-**B**ased **A**ggregate **F**orecasting
CPD	**C**onditional **P**robability **D**istribution
DAG	**D**irected **A**cyclic **G**raph
EM	**E**xpectationâĂŞ**M**aximization algorithm
FP-Growth	**F**requent **P**atterns-Growth
FP-Tree	**F**requent **P**atterns-Tree
IA	**I**mmune **A**lgorithm
IR	**I**mbalance **R**atio
JPD	**J**oint **P**robability **D**istribution
KDD	**K**nowledge **D**iscovery in **D**atabases
Kulc	**Kulc**zynski measure
LR	**L**inear **R**egression
LWL	**L**ocally **W**eighted **L**earning
MLP	**M**ulti-**L**ayer **P**erceptron
OvA	**O**ne-**v**s-**A**ll
SOM	**S**elf-**O**rganizing **M**ap
SSE	**S**um of the **S**quared **E**rrors
SVM	**S**upport **V**ector **M**achine
SVR	**S**upport **V**ector **R**egression
UK-Dale	**UK** **D**omestic **A**ppliance **L**evel **E**lectricity dataset
UKERC-EDC	**UK** **E**nergy **R**esearch **C**entre **E**nergy **D**ata **C**entre

Abstract

The big data framework applied to smart meters offers an exception platform for data-driven forecasting and decision making to achieve sustainable energy efficiency. Buying-in consumer confidence through respecting occupants' energy consumption behavior and preferences towards improved participation in various energy programs is imperative but difficult to obtain. The key elements for understanding and predicting household energy consumption are activities occupants perform, appliances and the times that appliances are used, and inter-appliance dependencies. This information can be extracted from the context rich big data from smart meters, although this is challenging because: (1) it is not trivial to mine complex interdependencies between appliances from multiple concurrent data streams; (2) it is difficult to derive accurate relationships between interval based events, where multiple appliance usage persist; (3) continuous generation of the energy consumption data can trigger changes in appliance associations with time and appliances. To overcome these challenges, we propose an unsupervised progressive incremental data mining technique using frequent pattern mining (appliance-appliance associations) and cluster analysis (appliance-time associations) coupled with a Bayesian network based prediction model. The proposed technique addresses the need to analyze temporal energy consumption patterns at the appliance level, which directly reflect consumers' behaviors and provide a basis for generalizing household energy models. Extensive experiments were performed on the model with real-world datasets and strong associations were discovered. The accuracy of the proposed model for predicting multiple appliances usage outperformed support vector machine during every stage while attaining accuracy of 81.65%, 85.90%, 89.58% for 25%, 50% and 75% of the training dataset size respectively. Moreover, accuracy results of 81.89%, 75.88%, 79.23%, 74.74%, and 72.81% were obtained for short-term (hours), and long-term (day, week, month, and season) energy consumption forecasts, respectively.

Chapter 1: Introduction

1.1 Introduction

Millions of homes worldwide are currently being equipped with smart meters, capable of generating data measurements for more than 100 energy consumption data points every 15 minutes resulting in a massive volume of data[1] that is smart meters big data. Smart meters offer bidirectional communication between consumers and utility companies, which has given rise to pervasive computing environments that generate extensive volume of data with high velocity and veracity attributes. Such data has a time-series notion typically consist of energy usage measurements of component appliances over a time interval [2][3]. The advent of big data technologies, capable of ingesting this large volumes of streaming data and facilitating data-driven decision making through transforming data into actionable insights, has revolutionized capabilities for how consumer energy usage decision patterns are learned, how energy demand are forecast, how outages are prevented, and how energy usage is optimized. Additionally, responsible, efficient and environmentally aware energy consumption behavior is becoming a necessity for reliable smart grid. Therefore, active participation by end-user/consumer, adopting such behaviors, can achieve substantial cost reductions in household energy bills [4].

Utility companies are consistently working to determine the best ways to reduce costs, improve profitability, and achieve sustainability by introducing programs, such as demand side management and demand response, that best fit consumers' energy consumption profiles. However, there has been marginal success in achieving the goals of such programs. Sustainable results are yet to be accomplished [5] because understanding

consumers' individual habits to tailor strategies that take into account both the benefits and limitations of modifying behavior according to suggested energy savings plans is difficult. The relationship between human behavior and the parameters affecting energy consumption patterns are non-static [5]. For example, consumer behavior is dependent on seasonal weather changes, which have variable influence over energy consumption decisions. Progressive learning of consumer behavioral energy consumption decision patterns is essential to support effective, well informed and time appropriate decision making at all levels, from the household to the system. Thus, actively engaging consumers in personalized energy management by facilitating well timed feedback on energy consumption and related costs can provide suitable support for energy saving strategies [6].

As advanced energy efficiency programs for households are introduced, such as automatic demand response, it is crucial to understand how individual temporal consumption and behavioral choices are reflected in consumption patterns to gain consumers' confidence for these programs. Traditionally, demand response mechanisms have been used by utility companies to target appliances that consume large amounts of power during peak times; these mechanisms reduce the mode of energy consumption or shift appliance operation to off-peak times. Currently, companies seek to gain acceptance for demand-response mechanisms among residential consumers because current demand response techniques pay little attention, if not at all, to usage correlation relationships with other appliances, resulting in users' preference requirements not being met.

It is important to design models that analyze and visualize energy consumption data from smart meters to uncover various temporal energy consumption patterns that directly reflect consumers' behaviors and expected comfort. These models must be able to extract consumer temporal energy consumption patterns at the appliance level

because these patterns define temporal appliance usage not only by the hour but also by what time of the day, period of day, week, month, or season usage patterns occur along with appliance-appliance associations. Understanding these appliance-appliance and appliance-time associations is essential to analyze the factors that impact consumers' behavior in relation to energy consumption, and capturing varying influence of consumers' energy consumption decision patterns to reflect most up-to-date energy needs via continuous online learning is vital for tese models [7]. Additionally, the models must be capable of predicting multiple appliance usage over a short and long term frame from appliance usage patterns, and forecast energy household energy consumption. Such predictions can serve as key parameter to the success of the smart grid energy saving programs in different ways. For example, daily energy predictions can be used to optimize scheduling and allocation. Weekly energy predictions can be used to plan energy purchasing policies and maintenance routines while monthly and yearly energy predictions can be used to balance the grid's production and strategic planning. Additionally, using these energy predictions consumers can cut-down their energy cost by actively engaging in energy management.

1.2 Research Motivation

The end use of energy in residential premises is based on the activities that occupants perform, the time at which appliances are used, and the interdependencies between appliances used simultaneously. For example, an user might charge an electric vehicle while the washing machine is on, might operate the dishwasher while the dryer is on, or work on the computer or watch television while cooking and listening to music. Additionally, time-of-use for appliances might differ between households; for example, at one

house washing clothes and using dishwasher might be done at night whereas in another house these activities might occur in the afternoon during the weekends. Thus, these relationships that is appliance-appliance (inter-appliance) and appliance-time associations have direct effect on behavioral traits or behavioral energy consumption decisions of consumers.

Due to the aforementioned heterogeneity and complexity of how consumers go about their energy usage decisions, understanding the consumers' energy usage behavior can be very costly: it is not feasible to contact every consumer and obtain their energy consumption characteristics. However, massive amounts of data are continuously being collected through household smart meters, which provides a unique opportunity for developing analysis models to help consumers and utility companies realize and unlock the potential benefits of investing in smart grid energy-saving programs. Smart meters big data as a large time series has a wealth of information about consumers' energy consumption behavior and contain many frequent patterns of appliance usage [3], which is becoming of interest to energy producers, utilities and end-users to mine such patterns from the ever growing data [1].

The study presented in this thesis was primarily motivated by non-trivial nature of mining energy consumption decision patterns from smart meters data, need to efficiently represent the occupants' behavioral traits and preferences, and determine methods to incorporate these patterns into energy program to achieve energy efficiency. Additionally, the rich content of energy consumption data and potential to extraction of this valuable information for many decision-making processes, such as consumer energy consumption behavior analysis, demand response optimization, and the improvements in

energy reduction recommendations provided further encouragement to take up this research.

1.3 Research Problem

In recent times it has become apparent that household energy conservation is a significant challenge but offers lots of opportunities for all; i.e., energy producers, policymakers, and end users. Greenhouse gas emissions and climate change are two widely accepted after-effects that are raising serious concerns over current energy consumption practices and emphasizing the need for sustainable energy behavior. Therefore, responsible, efficient and environmentally aware energy consumption behavior is becoming a necessity for reliable cyber physical systems such as the smart grid. There is an unswerving influence of the human behavior on energy consumption patterns at household that has direct impact on household energy consumption, which can be learned and taken into account by analyzing appliance level usage and appliance associations with time and other concurrently used appliances. Thus, utilities, as well as end users of energy, require mechanisms to seamlessly analyze end-user consumption behavior to make informed decisions about smart grid energy saving programs. Additionally, this decision making process must base on most recent, updated and time-appropriate information. The rise of smart meters big data has facilitated machine learning, predictive analysis, and data visualization for data-driven forecasting and decision making to improve energy programs, but mining and determining consumer behavioral implications for household energy consumption is complex and not trivial.

The research problem of this thesis focuses on a number of technical issues related to discovering, analyzing and visualizing consumer energy consumption behavioral patterns. These issues can be summarized as follows:

First, the influence of human behavioral traits on household energy consumption must be dynamically captured, both quantitatively and qualitatively, based on household energy consumption in the form of inter-appliance and appliance-time associations, which can act as one key criterion to energy savings and related programs.

Second, besides identifying appliances responsible for peak power load, it is required to identify the appliances (both manual and automatic operation) responsible for substantial contributions toward household energy usage to support active and efficient energy management at the household level. But, it is challenging to recognize such appliances without analyzing raw energy consumption data.

Third, considering the continuous nature of data generation from smart meters, an online incremental progressive data mining mechanism must be devised to learn varying appliance associations while capturing occupants' behavioral deviations through energy consumption patterns, which represent of occupants' energy consumption preferences and related anticipated level of comfort in relation to the use of electrical appliances. Addressing this issue will allow energy programs to obtain consumer confidence and participation towards greater success. However, the primary challenge here is how to derive accurate relationships between interval-based events during which multiple concurrent appliance usage persists.

Finally, real-world scenarios, which have multiple appliances operating concurrently, result in multiple concurrent time series. This would require a mechanisms to

predict multiple appliances concurrent usage to build foundation for household energy forecast at short-term and long-term time frames.

1.4 Research Approach

The proposed incremental mining and prediction model provides a framework to measure and analyze energy usage changes caused by consumer behavior. The model needs an online, dynamic, and recursive mechanism to extract the information from the energy consumption data that is most immediate representation of consumers' energy consumption behavior. The data from the smart meters are recursively mined in quanta of 24 hours, and the results; i.e., discovered inter-appliance and appliance-time associations, are maintained across successive mining exercises. In other words, data mining can be viewed as a process being conducted at the end of each day in an incremental but progressive manner, where only a portion of the entire database is mined at each iteration; thus reducing the memory overhead and achieve improved efficiency. Therefore, making it suitable for the real-world online applications, where data generation is a continuous process. In this context, an unsupervised machine learning approach is most appropriate, which eliminates the need to re-learn by re-mining the entire database (database + incremental) at frequent intervals that can be very expensive in case of supervised machine learning strategy. Although, a supervised learning approach appears to get benefited from consumer inputs, but with the massive size of smart grid and enormous number of consumers involved in the process, it might not be practically feasible to collect such input preferences from all the consumers at regular intervals. Moreover, these consumer preferences might change during the data mining operations and produce incorrect results.

Due to perpetual nature of production of energy consumption data by smart meters, over a period of time the inter-appliance associations can change and/or new ones can establish, which is indicative of consumers' energy consumption decisions. The proposed mechanism captures these variations comprehensively. Additionally, a Bayesian network based probabilistic graphical model is incorporated to predict multiple appliances usage and to forecast household energy consumption. The proposed model is capable of short-term predictions ranging from hourly to 24 hours intervals, and long-term predictions for days, weeks, months, or seasons. To evaluate the proposed mechanism three datasets were used: (1) the UK Domestic Appliance Level Electricity dataset (UK-Dale) was used [8], which is a time series data of power consumption collected between 2012 and 2015 with a time resolution of six seconds for five houses containing 109 appliances in Southern England; (2) Almanac of Minutely Power dataset (AMPds2) [9], which is a time series data of power consumption collected from a residential house in Canada between 2012 and 2014 at a time resolution of one minute; and (3) A synthetic dataset containing energy consumption measurements for one house with one minute resolution for a year.

1.5 Main Contributions

Comprehensive consideration of human behavioral variations related to energy consumption including high uncertainty in the order of use, varying time of use, and increased or reduced frequency of use of appliances, is necessary for supporting data-driven, well-informed, and time-appropriate decision making. The main contributions of this thesis are as follows:

- Behavioral energy-consumption pattern mining for appliance-appliance and appliance-time associations, derived from smart meters data, is proposed to provide insight into consumers' energy consumption decision patterns. The Appliances of Interest (AoI), which usually have smaller power load footprints but are major energy consumers due to extensive use, were determined. This information is important for informing consumers who tend to monitor appliances that consume large amounts of energy for short periods of time (e.g., washing machines or dryers) but overlook small appliances that contribute small energy footprints for longer duration. This is vital to achieve efficient energy demand management.

- Incremental progressive data mining is proposed as a suitable method for capturing consumer behavioral variations, learned through appliance associations, in the smart grid environment. Both FP-growth and k-mean algorithms were extended to incorporate incremental behaviors suitable to mine smart meters energy consumption data at appliance level. This model mines data in an online and distributed fashion at the individual household level to support energy efficiency and improved accuracy for decision making. The k-means clustering through dynamic programming, dynamic determination of k (number of clusters) through use of $Silhouette\ coefficient$ for k-means clustering, and application of Kulczynski measure ($Kulc$) along with imbalance ratio (IR), in frequent pattern mining, as pattern interestingness measures to eliminate uninteresting patterns and rules were used.

- A Bayesian network based probabilistic model for energy consumption prediction was utilized to predict all possible appliances expected to operate concurrently. The results were used to forecast household energy consumption for short term and long term time frames. The other applications of multiple appliances prediction could

include determining energy consumption behavioral patterns, and predicting daily consumer activity to support smart grid energy efficiency programs.

1.6 Publications

1.6.1 Conference Papers

1. **Shailendra Singh** and Abdulsalam Yassine and Shervin Shirmohammadi, *"Incremental Mining of Frequent Power Consumption Patterns from Smart Meters Big Data"* in IEEE Electrical Power and Energy Conference 2016.

1.6.2 Journal Papers

1. **Shailendra Singh** and Abdulsalam Yassine and Shervin Shirmohammadi, *"Mining Smart Meters Big Data for Behavioral Analytics and Energy Consumption Prediction"* in IEEE Transactions on Big Data (TBD) : Cyber-Physical Systems 2016. (Submitted)
2. **Shailendra Singh** and Abdulsalam Yassine and Shervin Shirmohammadi, *"Households Energy Consumption Patterns Mining for Effective Demand Response in Smart Grids"* in IEEE Transactions on Emerging Topics in Computing (TETC) : Big Data Computing for the Smart Grid 2016. (Submitted)

1.7 Thesis Organization

The organization of thesis is as follows:

Chapter 1: Introduction :: The thesis, research motivation, related problems, research goals, contributions of the research, and publications are discussed.

Chapter 2: Background and Related Work :: Fundamental concepts of smart grids, and knowledge discovery and data mining (KDD) associated with this research, and literature review of recent and related research are covered.

Chapter 3: Progressive Incremental Data Mining - Proposed Model :: The architecture of the proposed model with respective building blocks for the various phases such as frequent pattern mining, cluster analysis, predicting usage of multiple appliances, forecasting energy consumption, and visualization are presented.

Chapter 4: Progressive Incremental Data Mining - Frequent Patterns Mining :: Frequent pattern mining to discover appliance-appliance associations and association rules using an extended FP-growth strategy and Apriori approach with related results are described and discussed.

Chapter 5: Progressive Incremental Data Mining - Cluster Analysis :: Cluster analysis to discover appliance-time associations including the hour of day, time of day, weekday, week, month and season along with related results described and discussed .

Chapter 6: Multiple Appliance Usage Prediction and Energy Consumption Forecast :: A probabilistic model based on a Bayesian network, which utilizes historical evidence learned from frequent pattern mining and cluster analysis, is explained along with multiple appliance usage prediction and energy consumption forecast results.

Chapter 7: Conclusion and Future Work :: A conclusion of the research undertaken and

future work indicating direction for subsequent research is presented.

Chapter 2: Background and Related Work

2.1 Background

2.1.1 Smart Grids

A smart grid is an electrical grid capable of two-way communication, comprising of electricity generation stations, distribution and utility networks, renewable energy resources, and residential and commercial consumers equipped with advanced metering infrastructure (AMI) to facilitate efficient energy management. Smart grids offer several benefits such as reliability, load balancing, peak leveling, sustainability, demand-response management, and integration of multiple sources of power generation, including solar, and wind along with conventional modes [10] [11] [12]. In a smart home, many of the smart appliances can be connected together over a wireless network creating a context aware environment where energy consumption can be measured at appliances level. This can facilitate computerized control of appliances and home to appropriately respond to signals from energy providers or utilities and enable effective energy management. At the same time traditional homes and appliances can be equipped with low cost add-on infrastructure such as smart outlets, smart light switches, and communication hubs to make them smart and exploit full strength of smart grid environment and extract maximum possible benefits. Figure (2.1) by [13] presents the typical architecture of a smart grid. In a smart grid, data generation takes place at multiple points, but this research focuses on consumption of electricity to analyze massive amount of data generated in form of energy consumption patterns. Electricity consumption is conducted by end-user or consumer in homes, commercial buildings, and electric vehicle charging stations. Here, residential

homes and how end-users consume energy are considered to study the impact of consumers' behavior on household electricity consumption and to develop a methodology for predicting energy consumption using data mining.

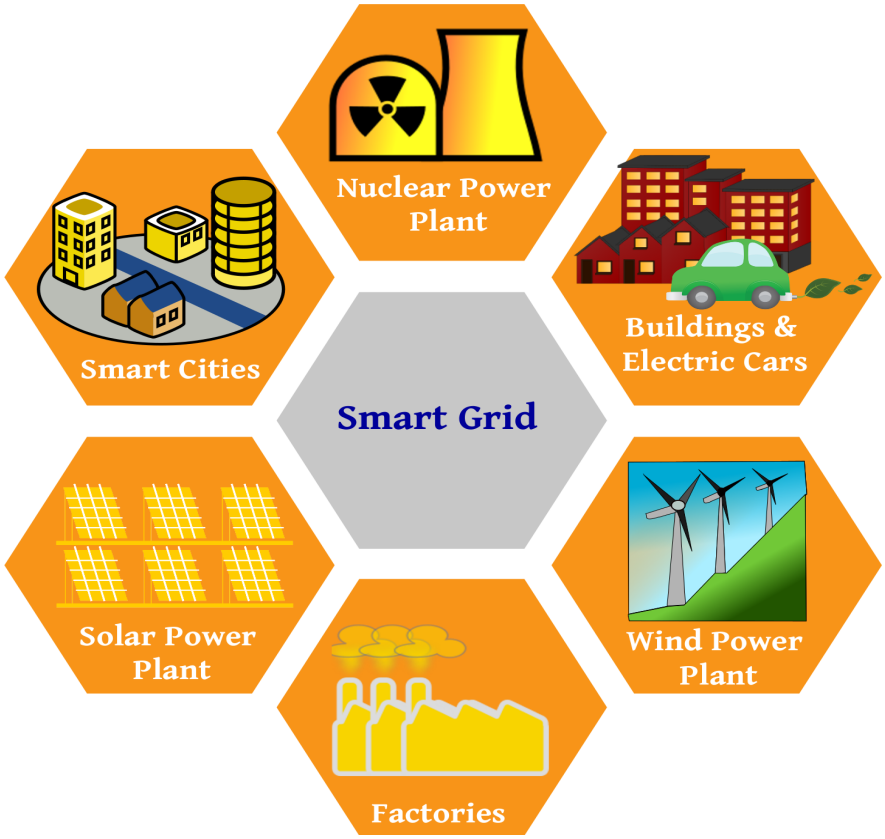


FIGURE 2.1: Smart Grid [13]

2.1.2 KDD - Knowledge Discovery in Databases

KDD is the process of extracting critical but relevant and valuable information from a database. It contains five steps, as depicted in figure (2.2) by [14], to process raw data and harvest information. KDD is an iterative process requiring constant user interaction. A

prerequisite step is to obtain the application domain knowledge, which enables the selection of the most appropriate dataset to unearth the information. Next, the data are cleaned by eliminating noise or outliers and inconsistent data; this accompanied by extracting details on outliers, missing, unknown, or incomplete data along with mapping with appropriate replacements. During this step, data from multiple sources are also combined by conducting data integration. Then, data reduction, projection and transformation is performed to reduce data dimensions by selecting the required data features. Next, data mining strategy is identified based on the best fit with the application or task under consideration and data mining is performed to reveal patterns. Finally, data mining outcomes (i.e. patterns) are evaluated for usefulness based on selected interestingness measures and irrelevant patterns are eliminated; but valuable patterns are translated and visualized in a ready to ingest form [14] [15].

Data mining, which is one key step in the KDD process, can be predictive, descriptive or both and is used to search and extract useful patterns of interest from enormous volumes of data. These patterns can be association rules, classification trees, and/or clusters which are discovered using specific data mining algorithms [14] [16] [17]. In this research, frequent pattern mining and cluster analysis were used to discover appliance-appliance and appliance-time associations related to energy consumption patterns respectively. This information was used to create an unsupervised probabilistic predictive model by combining the Bayesian network with cluster analysis and frequent pattern mining to predict multiple appliance usage and forecast household energy consumption on short term and long term basis.

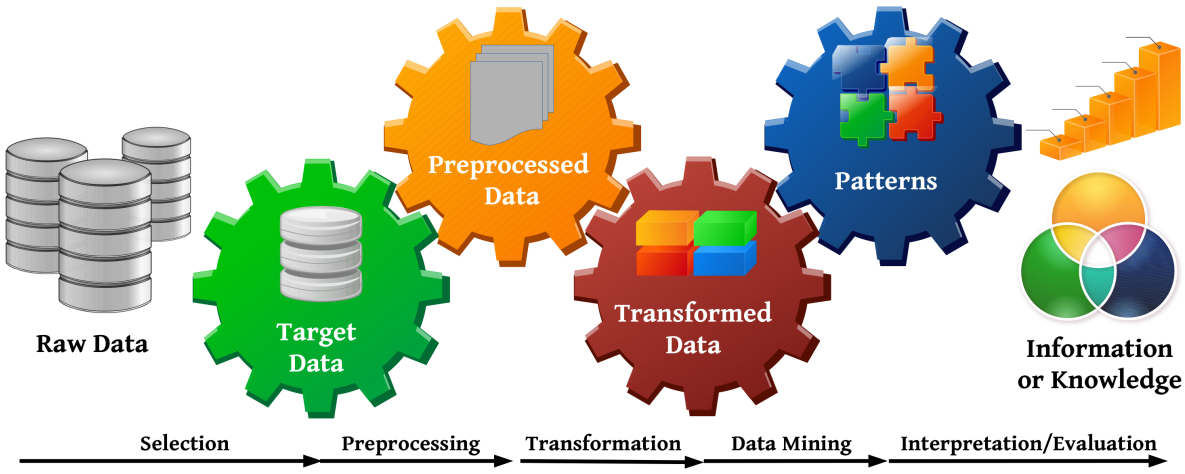


FIGURE 2.2: KDD - Knowledge Discovery in Databases : Process [14]

Frequent Pattern Mining

A frequent pattern is a set of items or sub-sequences that occur regularly in a dataset. Frequent pattern mining searches for these recurring patterns in a given dataset to determine associations and correlations among patterns of interest [18]. This research incorporated both, the FP-growth mechanism and the Apriori algorithm to discover appliance-appliance associations in the form of frequent patterns and association rules, respectively. Frequent pattern mining and its results are presented in Chapter (4).

Cluster Analysis

Cluster analysis is the process of grouping data objects into classes, which exhibit similarity to each other but dissimilarity with objects in other classes or clusters. This is an unsupervised form of classification which is capable of distinguishing groups or classes of objects and where classes are learned from the data [18]. There are various clustering approaches such as hierarchical clustering, centroid-based or partitional distribution

clustering, distribution-based clustering, and density-based clustering. Here, in this research k-means, a partitional distribution clustering algorithm was extended to discover appliance-time associations, which is described in Chapter (5).

Bayesian Network

A Bayesian network is a probabilistic graphical model consisting of random variables and their dependencies represented by nodes and arcs on a Directed Acyclic Graph (DAG). Each node in a probabilistic network is associated with a conditional probability table specifying its Conditional Probability Distribution (CPD) and facilitates the computation of Joint Probability Distributions (JPD) for the model [19] [18]. The model was formulated by learning historical evidence through cluster analysis and frequent pattern mining in form of appliance-appliance and appliance-time association respectively, to predict usage of multiple appliances concurrently and energy usage, as discussed in Chapter (6).

2.2 Related Work

Learning occupants' behavioral characteristics towards energy consumption is one key to the success of energy saving programs. Behavioral analytics as an approach to understand and predict appliance usage and energy consumption is relatively new. Technical report by [5] provides an extensive argument in support of exploiting behavioral energy consumption information to encourage and obtain greater energy efficiency. Mining appliance usage and association in form of frequent patterns (inter-appliance) and clusters (appliance-time) from context-aware smart-meter data can reveal surprising underlying

information. In this section, we review existing work in the literature, which employ energy consumption data of smart meters to analyze consumers' energy usage behavior.

2.2.1 Behavior Analytics

Technical report by [20] discussed the need to analyze the behavioral characteristics to recognize solutions to empower the consumers to regulate the energy consumption towards reducing expenses. Extensive argument in support of exploiting the behavioral energy consumption information to encourage and obtain greater energy efficiency was made in [5] and [6]. Authors in [21] studied prospective to extract information about the occupants from the energy consumption patterns. The need to learn the human behavior changes through the context-aware sensing to enable the humans with decision making capabilities at required time is studied in [22]. The impact of the behavioral changes for energy savings were also examined by [23] and [24] and end-user participation towards the effective and improved energy savings were emphasized. Prediction of consumers' energy consumption behavior is also studied in several papers. For examples, the work presented in [25] uses a Bayesian network to predict the occupant behavior as a service request using a single appliance, but does not provide a model to be applied for real-world scenarios. The study by [26] inspects the rule-mining based approach to examine the related behavioral characteristics and identify association between the energy consumption and the time of appliance usage to assist energy conservation, demand-response and anomaly detection, but lacks a formal rule mining mechanism and fails to consider appliance association of greater degree. The authors in [27] suggested a clustering approach to identify the distribution of consumers' temporal consumption patterns, however, the

study does not consider the appliance level usage details, which are a direct reflection of consumers' comfort and does not provide correlation between generated rules and energy consumption characterization.

The approach provided by [28] suggests an auto regression model to compute the energy consumption profiles for residential consumers to facilitate the energy saving recommendations without the consideration of occupants' behavioral attributes. Study [29] proposed to map the occupants' activities responsible for the energy household energy consumption and suggest improved energy efficient ways to perform daily activities. The work in [30] proposed graphical model based algorithm to prediction the human behavior and inter-dependency for appliance usage activities to predict multiple appliance usage using a Bayesian model, but do not consider occupants' behavioral variations rather assume behavior is a cyclic and stationary phenomenon. Study [31] analyzed the Hidden Markov Model and cluster analyses to establish correlation between the energy consumption patterns and the user behavioral characteristics to support energy program enrollments. In the work [4] and [32], authors utilized the historical consumption behavior to decide appliance scheduling for the efficient energy usage and emphasize on the active participation by end-user/consumer to achieve significant reduction in household energy consumption; but failed to consider inter-appliance associations. Study [33] proposed methodology to extract user behavioral characteristics in form of the Activities of Daily Living (ADL) tasks with context and temporal information to support efficient power management.

Authors in [34] used a Semi Markov Model to detect the occupants' habits with a target to identify ADL adapting to the human behavior. Study in [35] utilized an electrical appliance usage signatures or energy consumption patterns to identify the occupants in

a house. In study [36], a log Gaussian Cox process is used to identify the unique daily routine to identify and monitor occupants' behavior. In the work [37], personal appliance usage habits were learned from the appliance usage patterns, which are dependent on ADLs. Papers [38] and [39] present a system to mine the appliance usage patterns to determine the energy consumption behavior with an aim to conserve power. Research by [40] proposed an algorithm to reveal occupants' activity patterns from energy consumption data. Work in [41] used visualization method to extract the energy usage patterns from everyday activity patterns and provide insight into the electricity usage to support targeted campaigns. In the study [42], critical patterns from energy consumption data were mined to determine the activity of daily living profile, and short-term and long-term inconsistencies were utilized to detect health conditions of the occupants towards facilitating home health care.

Studies [43] and [44] presented a demand side management technique to control the peak load hikes by shifting high power load appliances from operating while minimizing discomfort to the occupants, but does not completely consider the behavioral variations that may occur due to the time such as weekdays vs weekends, months and seasons. In the study [45] authors proposed to analyze the impact of price variations on occupants energy consumption behavior across the houses. Researchers in study [46] used neural networks to study, recognize and monitor variations in the human behavior extracted from smart meter time-series data, as energy consumption profiles for individuals, data towards health monitoring. Work by [47] studied the energy consumption data to determine the groups of customers based on energy usage behavior and related variability. Paper [48] used frequent pattern mining to determine the current activity, and reveal irregularities in the power consumption behavior. Study by [49] proposed methodology

to learn and predict the occupant behavior using Support Vector Machine (SVM), C4.5, and Locally Weighted Learning (LWL) in an office environment.

2.2.2 Frequent Pattern Mining

Authors in [50] and [51], modeled a time-series multi-label classifier to develop the decision tree taking appliance correlation into consideration to predict the appliance usage, but considered only last 24 hour window for the future predictions along with appliance sequential relationships. The study in [26] inspected the rule-mining based approach to examine the related behavioral characteristics and identify the association between the energy consumption and the time of appliance usage to assist energy conservation, demand-response and anomaly detection, but lacks a formal rule mining mechanism and fails to consider greater degree of association among appliances. The work in [2] proposed a new algorithm to consider the incremental generation of data and mining appliance associations incrementally. Similarly in [52], appliance association and sequential rule mining was studied to generate and define the energy demand patterns.

The work presented by [3] and [53] mine for sequential patterns to understand the appliance usage patterns to save energy. Authors in [54] proposed an algorithm to mine probabilistic correlation patterns among the appliances; which in the work [55] was extended for incremental sequential mining technique to discover correlation patterns among appliances, where authors proposed new algorithm which offers reduction in memory with improved performance. Authors in [56] further extend their approach to

mine time interval data for probabilistic temporal pattern mining to discover the appliance usage patterns. Context aware association mining through frequent pattern recognition was studied in [57] where the aim is to discover the consumption and the operation patterns and effectively regulate the power consumption to save energy. Study by [58] proposed pattern recognition technique to identify ADLs from electrical appliance, but used only one appliance; which is extended in [59] to identify five concurrent operating electrical appliances using the real power consumption of appliances.

Authors in [60] used sequential pattern (activity) mining to predict the future activities enabling enhanced power consumption forecast, and present short term load forecasting using activity sequence patterns with Support Vector Regression (SVR). Paper [48] presented frequent pattern mining using Apriori approach to determine the current activity along with occupants' location in the house; proposed model is capable of revealing irregularities in the power consumption behavior.

2.2.3 Cluster Analysis

The authors in [27] suggested a clustering approach to identify the distribution of consumers' temporal consumption patterns, however, the study does not consider appliance level usage details, which are direct reflections of consumers' comfort and did not provide the correlation between generated rules and the energy consumption characterization. The study in [61] used clustering as a means to group the customers according to load consumption patterns to improvise on the load forecasting at a system level. Similar to [61], the authors in [62] used k-means clustering to discover the consumers' segmentation and used socio-economic inputs with SVM to predict the load profiles towards demand

side management planning. The work in [63] proposed a methodology to disclose the usage pattern using hierarchical and c-means clustering, multidimensional scaling, grade data analysis and sequential association rule mining; while considering appliances' ON and OFF events. However, the study does not consider the duration of appliance usage or the expected variations in the sequence of appliance usage.

The study in [64] employed hierarchical clustering, association analysis, decision tree and SVM to support short-term load forecasting, but does not consider variable behavioral traits of the occupants. The methodology proposed by [65] used a two-step clustering process to examine the load shapes and proposed a segmentation schemes with appropriate selection strategies for the energy programs and related pricing. [4], utilized historical energy consumption behavior to determine the appliance scheduling and computes the load profiles based on clustering of appliances having similar schedules to taper peak load and keep it as close as possible to the average load, but failed to consider inter-appliance associations. Work by [39] presented algorithms based on hierarchical clustering to mine the appliance usage patterns to conserve energy. Authors in [66] used the k-means and artificial neural networks to cluster customer profiles which can be used for the energy management. The study [67] suggested to use k-means clustering to determine the consumption categories and estimate the peak loads for the houses through the polynomial regression.

Study [68] proposed k-means along with Self-Organizing Map (SOM) to compute the baseline load estimate by clustering similar load patterns for demand response purpose. Study [69] used k-means in conjunction with Silhouette analysis towards the load profiling to design time of use tariffs. In the study [45] authors proposed to use fuzzy c-means clustering to compute daily load profiles. Work [69] proposed to utilize

k-means along with a polynomial regression technique to estimate the peak load for the traditional electrical meters, and uses global silhouette coefficient to determine the value of k. Authors in their study [47] used expectationâĂŞmaximization (EM) algorithm, along with Bayesian Information Criterion (BIC) to determine the number of clusters, to determine the groups of customers based on energy usage behavior. Study [70] utilized self organizing maps strategy of clustering for the daily peak load forecasting. In the study, [71] unsupervised Bayesian Clustering by Dynamics (BCD) classifier to detect the state of input load time-series and create sub-groups is used, which are further fed to SVR for forecast electrical load.

2.2.4 Multiple Appliance Usage Prediction and Energy Consumption Forecast

Researches in [72] proposed neural network based on Bayesian learning methods for short term load forecasting and demonstrate benefits over the classic neural network learning approach; where [73] suggested a Bayesian learning approach for artificial neural network (ANN) and provide evidence for better forecasting results while compare to the classical ANN. Study in [74] used ANN for the load forecast for several hours in future. Paper [70] utilized hybrid neural network model for the daily peak load forecasting using the self organizing maps strategy of clustering. In the study, [71] use of BCD and SVR is proposed for the electrical load forecast one day in advance.

In the study [75] used SVM for monthly electrical load forecasting. Research

[76] exploited SVR along with Immune algorithm (IA) for the electrical load forecast. Researchers in [77] evaluates various forecasting strategies Linear Regression (LR), Multi-Layer Perceptron (MLP), SVR, and propose algorithm Cluster-based Aggregate Forecasting (CBAF) for forecasting electrical load for one to 24 hours period, and provide support for better performance of the CBAF. Authors in [78] predicted appliance power consumption based on the historical power consumption and develop appliance level profiling to forecast the appliance usage up-to 90 minutes in future. In study [60] presented short term load forecasting using the activity sequence patterns with Support Vector Regression. Study [29] proposed linear SVM model as the best approach to predict the energy consumption for a smart environment while comparing with the linear regression and non-linear SVM model. Paper [67] proposed to use k-means clustering along with polynomial regression to estimate the household peak load.

2.2.5 Discussion

In all the above discussed approaches, a few examined the impact of consumer behavior on energy consumption but do not consider the human behavioral variations comprehensively such as high uncertainty in order of use of appliances to complete an activity, or increased or reduced frequency of use of appliances and variations in time of use due to seasons or other personal factors. It is important to note that these variations have a direct effect on household energy consumption. These studies analyze data through models that are static in nature; i.e., models are defined to learn one time and act according to design with no provision for incremental learning. And, in case, it is needed to process the incremental data generated by smart meters, these models would require

to re-training on entire dataset repetitively to reflect updated information, which will be very expensive. Additionally, appliance level usage and appliance associations are not fully considered, which can provide a necessary ground for effectively defining consumer energy consumption behavior with respect to which appliances are used, which appliances are used simultaneously, and when they are used. Therefore, efficient capturing of consumer behavior and related variations at appliance level can support accurate estimation and prediction of energy consumption. This, not only can assist utility companies customizing the energy programs to take consumer preferences into account for greater success through improved consumer engagement, but support consumers in their quest for effective energy management to reduce energy bills. This research addresses these shortcomings and improves further by adapting incremental progressive unsupervised machine learning through frequent pattern mining and cluster analysis. Thus, translating energy consumption patterns into frequent patterns and clusters representing inter-appliance and appliance-time associations, which are indicative of consumer behavior and related preferences, while explaining home energy consumption. Furthermore, these associations establish a probabilistic foundation (i.e. historical evidence) for predictions and this research uses a Bayesian network to develop prediction model capable of predicting appliance usage and energy consumption for short-term and long-term time frame.

Chapter 3: Progressive Incremental Data Mining : Proposed Model

3.1 Proposed Model

Frequent patterns are repeated patterns or itemsets, which often appear in a dataset. Within smart-meter data an itemset, for example, could comprise of laptop and washing machine that often presents themselves together in a frequent pattern. Hence, frequent pattern mining can reveal association and correlation among appliances that is appliance-appliance or inter-appliance associations. In addition to learning inter-appliance association, it is of critical interest to understand when the appliances are used with respect to hour of day (00:00 - 23:59), time of day (Morning, Afternoon, Evening, Night), weekday, week, month and season (winter, summer/spring, fall). The underlying information from smart meters time series data facilitates discovery of appliance-time associations through clustering analyses of appliances over time. Clustering analyses is the process of creating classes (non-supervised classification), where members of a cluster exhibit similarity with one another, but display dissimilarity with members of other clusters. Using such analyses, inter-appliance and appliance-time associations can represent critical consumer energy consumption behavioral characteristics, and identify peak load and energy consumption hours to explain respective behavioral traits. This can establish expected levels comfort for the consumers. Additionally, it is of significance interest to identify the AoIs to determine major contributors to energy consumption within household, and to develop capabilities to predict multiple appliance usage times and energy consumption forecasts on short term and long term time frames.

The mining of frequent patterns and cluster analysis are generally considered as an off-line and expensive process for large databases. In real world applications transaction data generation is a continuous process, in which new transactions are generated and old transactions may become obsolete as time progress, therby invalidating existing frequent patterns and clusters or establishing new associations and groups. Therefore, an incremental and progressive update strategy for energy consumption data mining is imperative to take into account variations and updates and ensure that discovered frequent patterns and clusters are duly maintained with updated information. For example, an appliance such as a space-heater will generally be used during winter, and usage frequency during other seasons is reduced. As an effect a significant gain during winter but decrease during other seasons will be registered. Consequently, the space-heater should appear higher on the list of frequent patterns and association rules during the winter, but much lower on the list during the summer or spring. Similarly, appliance usage frequency affects the size or strength of clusters that is association with time will update. The objective of capturing theses variations can be achieved through progressive incremental data mining while eliminating the need to re-mine the entire database at regular intervals. For a large database, frequent pattern mining can be accomplished using pattern growth approach [79], [80], whereas cluster analysis can be achieved using k-means cluster analysis [81]. Both these approaches are extended to fit in context of online incremental mining strategy of progressive manner and presented in this thesis.

In this proposed approach, available data is recursively mined in quantum of 24 hours and the results;i.e., discovered inter-appliance (frequent patterns) and appliance-time (clusters) associations are maintained across successive mining exercises. In other words, data mining can be viewed as a process being conducted at the end of each day

in an incremental manner. During each consecutive mining operation existing frequent patterns and clusters are updated and/or new patterns and clusters are added to the persistent database in a progressive manner. Using this technique, only a portion of the entire database is mined during each iteration, thus reducing memory overhead and achieving improved efficiency for real-world online applications, where data generation is a continuous process and there is a need to ingest this data and extract meaningful relevant information to support time relevant continuous decision making at various levels.

Frequent patterns and clusters discovered from databases can be maintained in-memory using a hash table data-structure or stored in a off the memory Database Management System. The latter approach reduces memory requirement at the cost of marginal increase in processing time, whereas the former approach reduces processing time but requires more memory. Considering the smart meter environment, quicker processing time is of importance; however, persistence of the information discovered through days, months, seasons, or years is more vital to achieve useful and usable results for the future. Therefore, we prefer permanent storage using Database Management System over in-memory volatile storage.

The proposed model extracts critical information from data in the form of frequent patterns (appliance-appliance association and related association/correlation rules), and clusters (appliance-time associations) using unsupervised incremental and progressive data mining techniques. Frequent patterns and clusters are used to determine the probabilistic associations among appliances and with time, and to identify AoIs with respective probabilities. The model is completed with Bayesian network based probabilistic prediction methodology to predict concurrent multiple appliance usage and forecast household energy consumption for both short term and long term time frames. For the

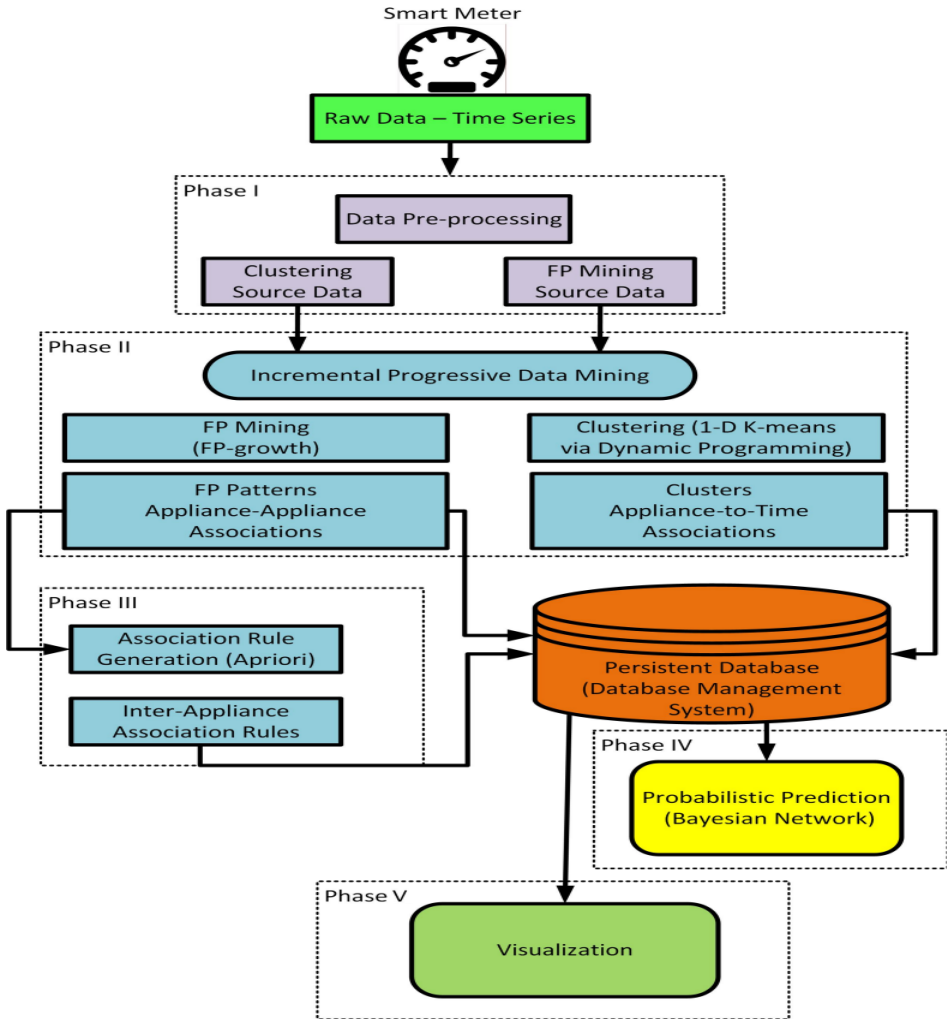


FIGURE 3.1: Model: Incremental Progressive Data Mining and Probabilistic Prediction

purpose of evaluation of the proposed model, the prediction results are then compared with a SVM based multiple appliance predictor outcomes for the accuracy of predictions. Why SVM? the choice is exclusively based on the acceptance of SVM, by research community, as one of the most capable prediction approach in the area of smart grid data analytics.

Figure (3.1) illustrates the proposed model with its distinct five phases: data

preparation/pre-processing, frequent pattern mining and cluster analysis, association rules generation, predictions, and result visualization. Brief descriptions of each phase is provided below, while detailed discussions with related theoretical background are presented in the respective chapters.

- Phase I: Raw data, which contains millions of records of energy consumption data from a smart meter, are prepared and processed for further analysis.
- Phase II: Incremental progressive frequent pattern mining and cluster analysis are performed. Frequent patterns and clusters are used to determine the probabilistic associations among appliances and with time, and to identify AoI.
- Phase III: Appliance-appliance association and correlation rules are extracted from appliance-appliance frequent patterns and associations.
- Phase IV: The Bayesian network based prediction mechanism ingesting the results from phase III to predict multiple appliance usage and energy consumption for short and long term time frames.
- Phase V: Visualization of the results.

3.2 Data-set and System Setup

Two real datasets were used for the research, UK-Dale dataset [8], which includes time series data of power consumption collected between 2012 to 2015 having a time resolution of 6 seconds for five houses with 109 appliances in Southern England published by UK Energy Research Centre Energy Data Centre (UKERC-EDC); and second, AMPds2 dataset [9], which is time series data for electricity, water, and natural gas measurements for over 2

years period from 2012 to 2014 having a time resolution of one minute for one residential house in the Greater Vancouver metropolitan area in British Columbia, Canada. In addition, a synthetic dataset is generated having over 1.2 million raw energy consumption data points with a time resolution of one minute from smart meter of one house with 21 appliances for a period of one year to conduct initial experiments. The underlying system for the proposed model was developed in Python, and the data storage used was MySQL and MongoDB databases on Ubuntu 14.04 LTS 64-bit system.

3.3 Data Preparation

Smart meters time-series raw data, which is a high time-resolution data, is transformed into one min resolution load data and subsequently translated into a 30 minutes time-resolution source data ;i.e., 24 * 2 = 48 readings per appliance per day, while recording usage duration, average load, and energy consumption for each active appliance. All the appliances that registered active during the a 30 minute time interval were included into the source database for frequent pattern data mining and cluster analysis. We analyzed and found time-resolution of 30 minutes as most suitable, because it not only captures

TABLE 3.1: Frequent Pattern Source Database

Start Time	End Time	Active Appliances
2013-08-01 07:00	2013-08-01 07:30	'2 3 4 12'
2013-08-01 07:30	2013-08-01 08:00	'3 4 12'
2013-08-01 08:00	2013-08-01 08:30	'2 4 12'
2013-08-01 08:30	2013-08-01 09:00	'4 12'
2013-08-01 09:00	2013-08-01 09:30	'2 3 12'
2013-08-01 09:30	2013-08-01 10:00	'2 3 4'
2013-08-01 10:00	2013-08-01 10:30	'2'
2013-08-01 10:30	2013-08-01 11:00	'12'
2013-08-01 11:00	2013-08-01 11:30	'2 12'
2013-08-01 11:30	2013-08-01 12:00	'3 12'
2013-08-01 12:30	2013-08-01 13:00	'2 4'
2 = Laptop, 3 = Monitor, 4 = Speakers, 12 = Washing Machine		

appliance-time and appliance-appliance associations adequately but also keeps the number of patterns to be analyzed sufficiently low and makes it appropriate for real-world applications. The actual dataset UK-Dale [8] had over 500 million raw records from five houses with a time resolution of 6 seconds. This was reduced to 20 million during preprocessing phase without loss of accuracy or precision. Similarly, dataset AMPds2 [9] was reduced to 4 million records from over 40 million raw records. Tables (3.1), (3.2), and

TABLE 3.2: Clustering Source Database - I

Appliance	Hour of Day	Time of Day
2	07:30 08:00	M E
3	16:00 16:30	A E
12	13:30 14:00 14:30	A
2 = Laptop, 3 = Monitor, 12 = Washing Machine		
M = Morning, A = Afternoon, E = Evening		

(3.3), show examples of the resulting ready to mine source data format comprising four appliances from one house.

TABLE 3.3: Clustering Source Database - II

Appliance	Weekday	Week	Month	Season
2	1 5	2 4	1 9	F W
3	3 5	3 5	8 10	F
12	2 4 6	1 3	1	W
2 = Laptop, 3 = Monitor, 12 = Washing Machine				
F = Fall, W = Winter				

3.3.1 Dataset : Raw Power Consumption Analysis

Extensive analysis was conducted on raw energy consumption data to examine power load and energy consumption trends for hour of the day, the time of the day, weekdays, months, and seasons. Figures (3.2), (3.3), (3.4), (3.5), and (3.6) show the average power load and (3.7), (3.8), (3.9), (3.10), and (3.11) show the average energy consumption trends

for sample results from one house; peaks for power load and energy consumption are noted for each time resolution.

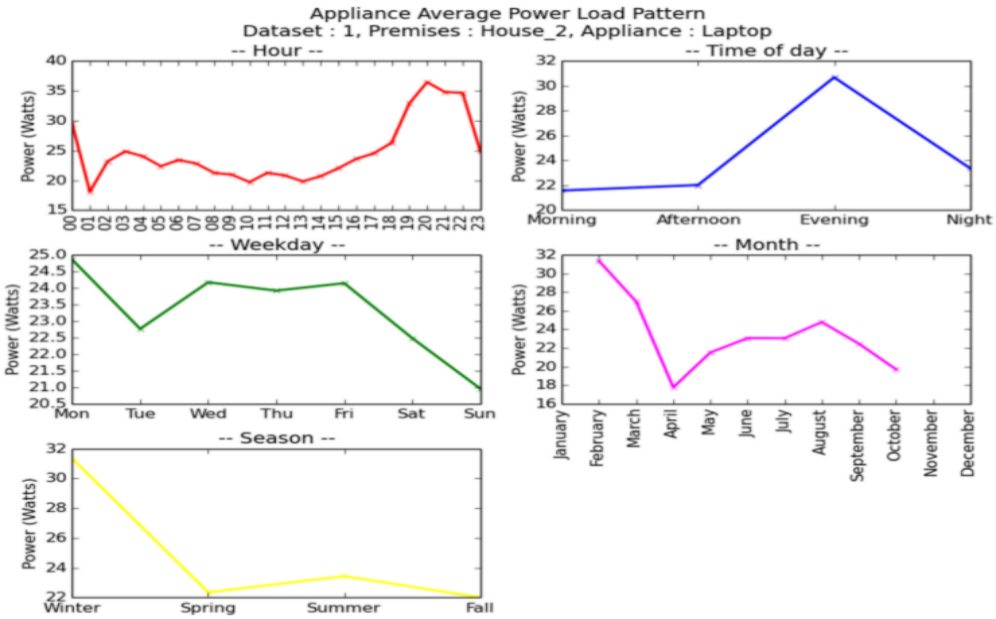


FIGURE 3.2: Dataset 1 House 2 : Average Power Load Pattern : Laptop

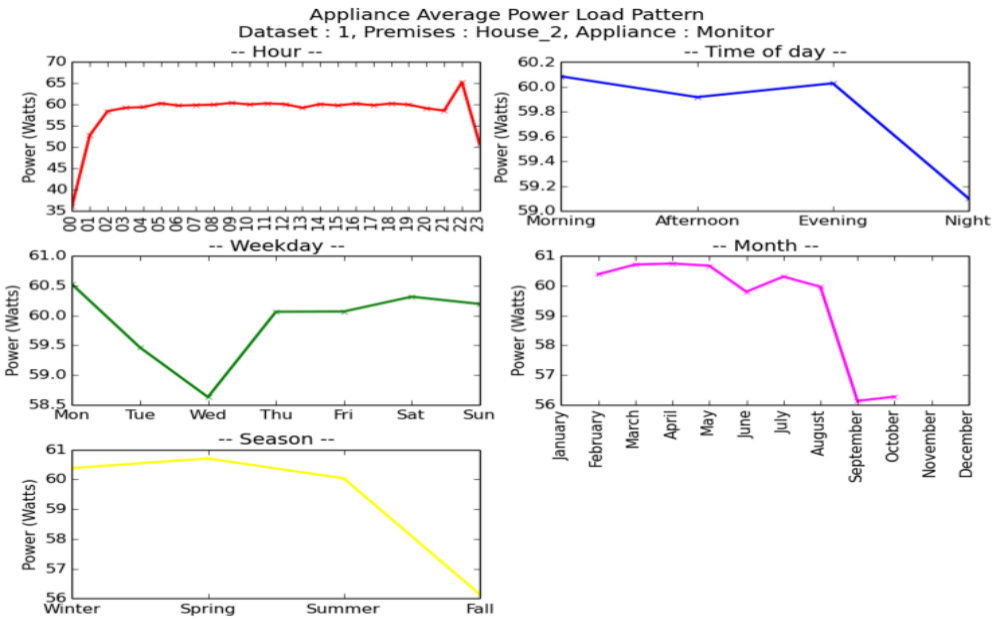


FIGURE 3.3: Dataset 1 House 2 : Average Power Load Pattern : Monitor

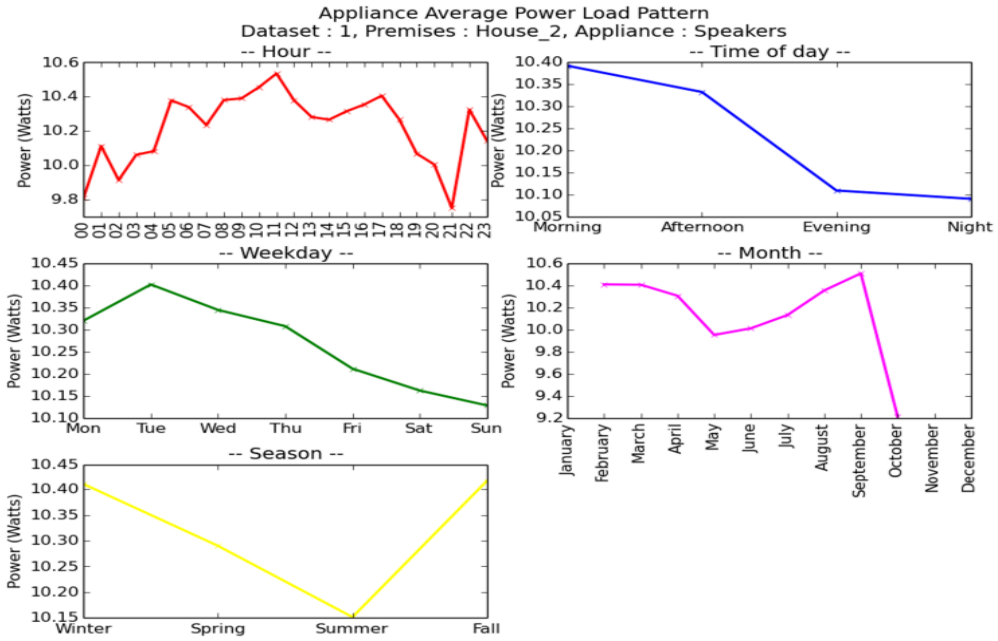


FIGURE 3.4: Dataset 1 House 2 : Average Power Load Pattern : Speakers

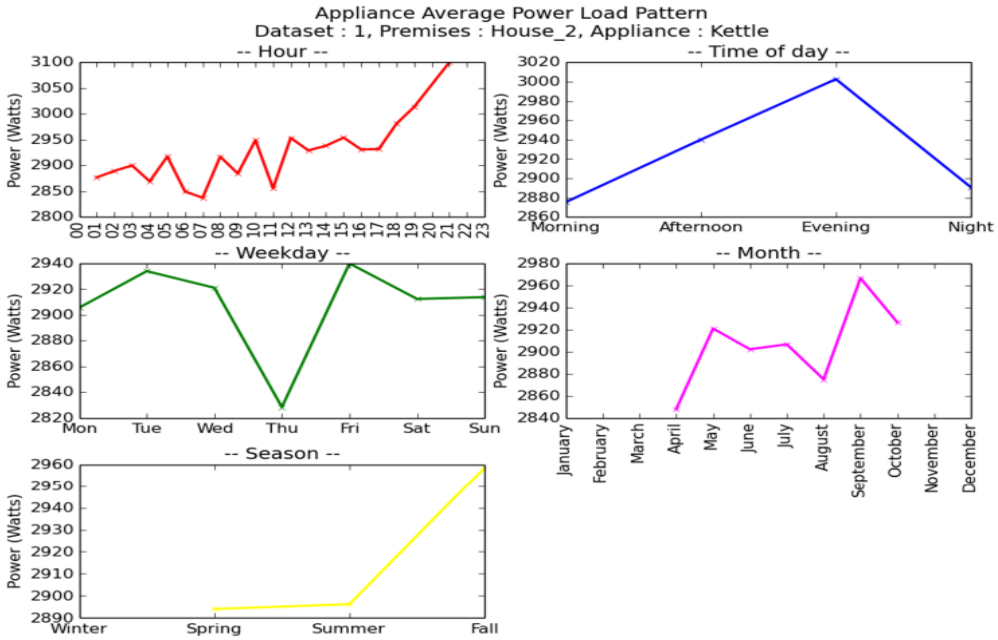


FIGURE 3.5: Dataset 1 House 2 : Average Power Load Pattern : Kettle

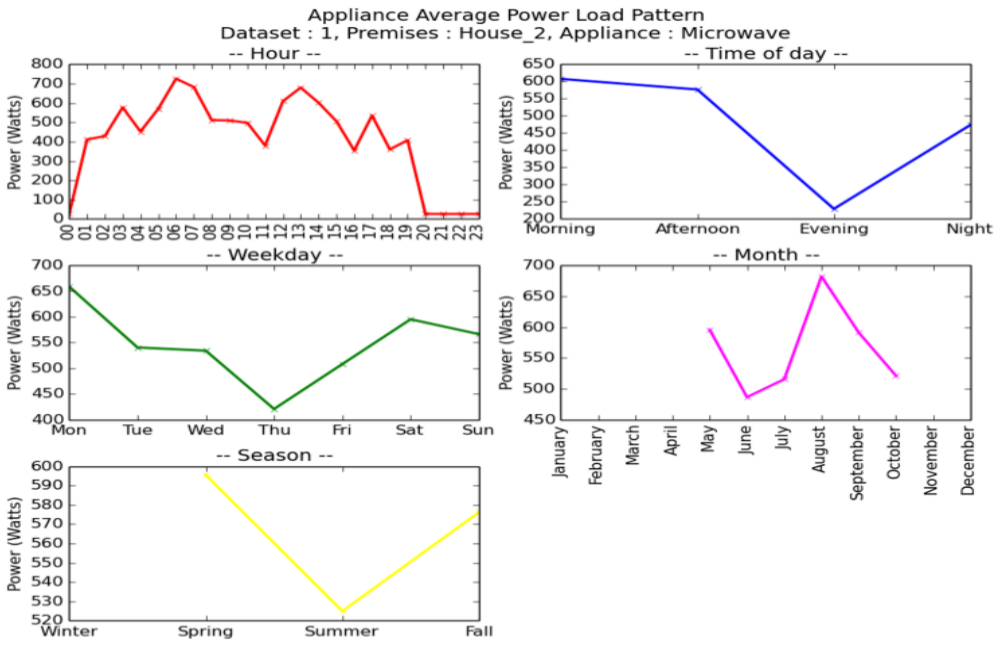


FIGURE 3.6: Dataset 1 House 2 : Average Power Load Pattern : Microwave

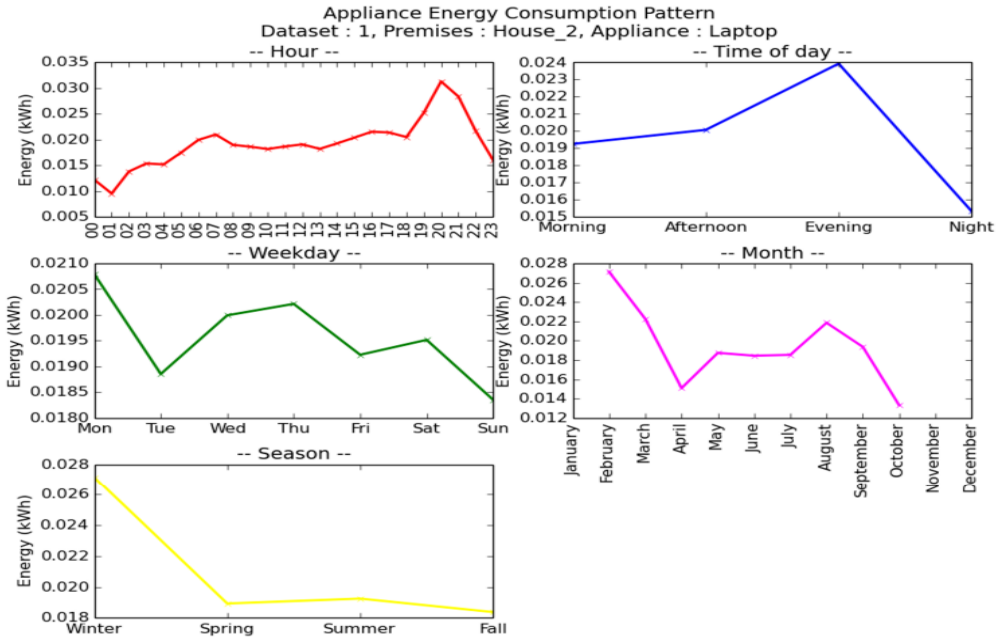


FIGURE 3.7: Dataset 1 House 2 : Average Energy Consumption Pattern : Laptop

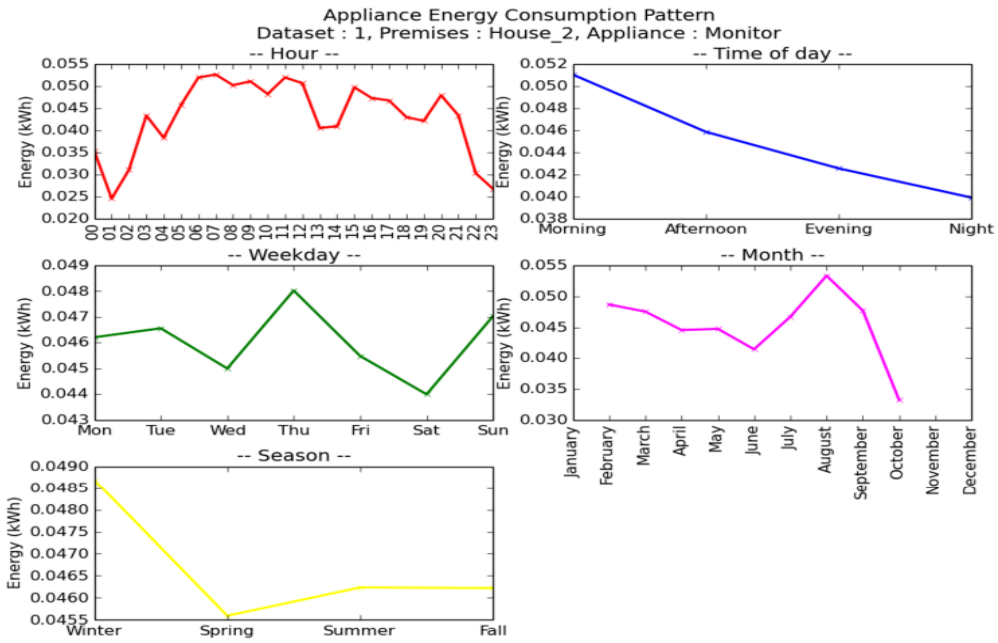


FIGURE 3.8: Dataset 1 House 2 : Average Energy Consumption Pattern : Monitor

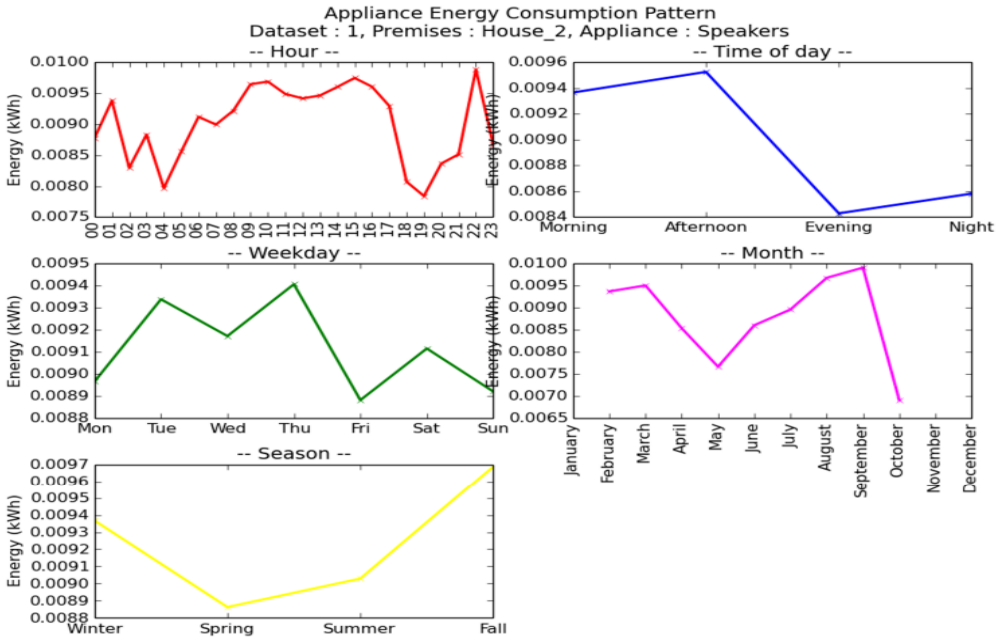


FIGURE 3.9: Dataset 1 House 2 : Average Energy Consumption Pattern : Speakers

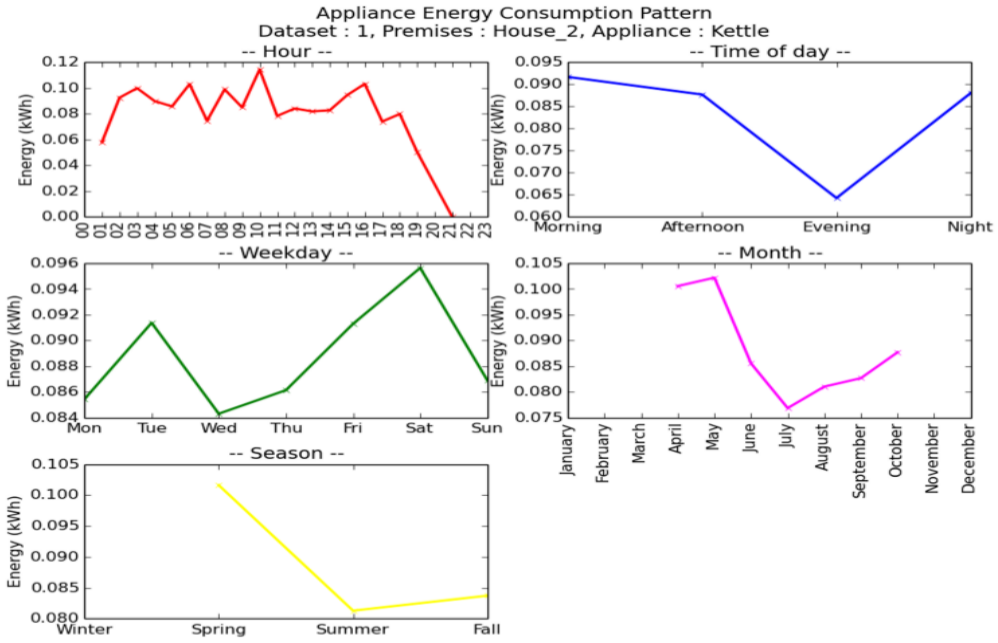


FIGURE 3.10: Dataset 1 House 2 : Average Energy Consumption Pattern : Kettle

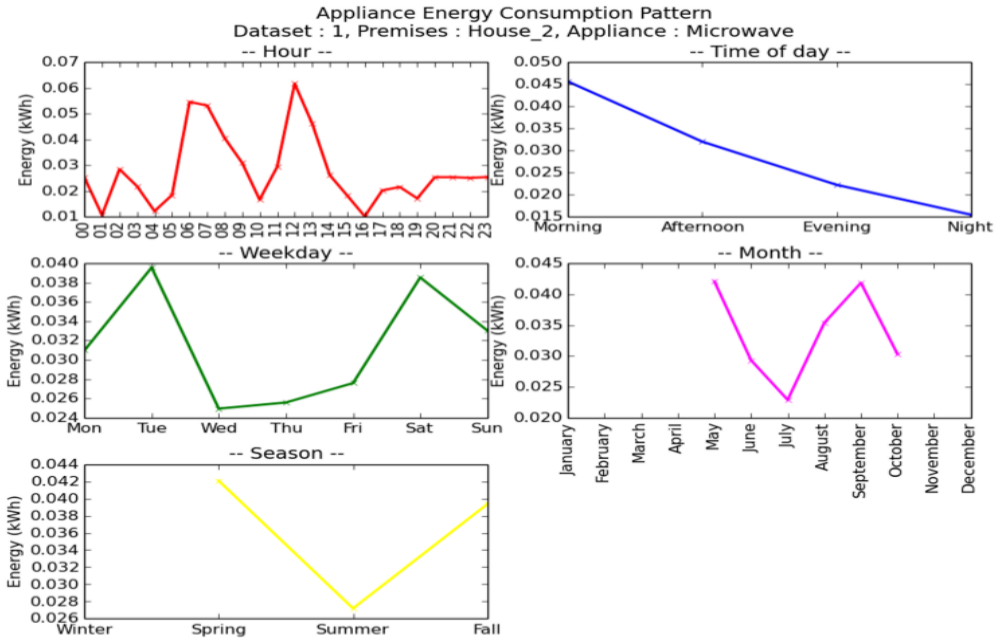


FIGURE 3.11: Dataset 1 House 2 : Average Energy Consumption Pattern : Microwave

Chapter 4: Progressive Incremental Data Mining: Frequent Patterns Mining

4.1 Frequent Pattern Mining

Frequent patterns are repeated patterns or itemsets, which often appear in a dataset. Considering smart-meter data, an itemset, for example comprising of laptop and washing machine, that often present itself together is a frequent pattern. Hence, frequent pattern mining can help discover association and/or correlation among appliances, which defines relationship among data interpreting consumer energy consumption behavior. Therefore, with the enormous quantity of data progressively being collected from smart meters, it is not only of keen interest to energy producers and utilities, but also to consumers to mine such frequent patterns for defining and facilitating decision-making processes such as energy saving plans, demand response optimization, and cost reduction.

4.2 Frequent Itemsets and Association Rules

In this subsection, preliminary background is introduced on frequent pattern mining based on [18]. Let $\Gamma = \{I_1, I_2, ..., I_k\}$ be an itemset containing k items which is referred to as k-itemset (l_k). Let DB, represent a transaction database with a set of transactions as described in table (3.1), where each transaction Υ is an itemset having $\Upsilon \subseteq \Gamma$ and $\Upsilon \neq \emptyset$. The frequency of appearance of an itemset is the number of transactions that contain the itemset, defined as the support count, or the count of the itemset. Let, X and Y be set

of items, such that $X \subseteq \Upsilon$ and $Y \subseteq \Upsilon$. Itemsets X and Y are considered frequent itemsets or patterns, if their respective $support$ s_X and s_Y that is the percentage of transactions the itemset appears in the transaction database DB, are greater than or equal to $minsup$; where $minsup$ is the pre-defined minimum support threshold. $support$ can be viewed as the probability of the itemset in the transaction database DB. This is referred to as the relative support, whereas the frequency of occurrence is known as the absolute support. Hence, if the relative support of an itemset $X(s_X)$ [or $Y(s_Y)$] satisfies a pre-defined minimum support threshold $minsup$, then the absolute support of X (or Y) satisfies the corresponding minimum support count threshold.

Association rules are the results of the second iteration of the frequent pattern mining process, where already discovered frequent itemsets or patterns are processed to generate the association rules. Rules, of the form $\{X \Rightarrow Y\}$, are generated using $support$-$confidence$ framework, where $support$ $s_{X \Rightarrow Y}$ [equation (4.1)] is the percentage of transactions containing $(X \cup Y)$ in transaction database DB, which also can be seen as the probability $P(X \cup Y)$. The $confidence$ $c_{X \Rightarrow Y}$ [equation (4.3)] is defined as the percentage of transactions in DB containing X that also contain Y, which is the conditional probability, $P(Y|X)$ [18]. Equations (4.1) and (4.3) captures the above notions respectively.

$$\begin{aligned} support(X \Rightarrow Y) = s_{X \Rightarrow Y} &= P(X \cup Y) \\ &= support(X \cup Y) \end{aligned} \tag{4.1}$$

$$absolute\_support(X \Rightarrow Y) = support\_count(X \cup Y) \tag{4.2}$$

$$\begin{aligned} confidence(X \Rightarrow Y) &= P(Y|X) \\ &= \frac{support(X \cup Y)}{support(X)} \\ &= \frac{support\_count(X \cup Y)}{support\_count(X)} \end{aligned} \tag{4.3}$$

Hence, an association rule established as $\{X \Rightarrow Y\}$, where $X \subset \Gamma$, $Y \subset \Gamma$, $X \cap Y = \emptyset$, $X \neq \emptyset$, and $Y \neq \emptyset$, with *support* $s_{X \Rightarrow Y} \geq minsup$ and $confidence$ $c_{X \Rightarrow Y} \geq minconf$ are classified as strong, where the $minconf$ is pre-defined minimum confidence threshold. Additionally, the association rule's support $s_{X \Rightarrow Y}$ will automatically satisfy the minimum support threshold as the rules are essentially generated from frequent patterns X, and Y having respective *support* $s_X, s_Y \geq minsup$. Thus, once the *support* for X, Y, and $(X \cup Y)$ are determined, corresponding association rules $\{X \Rightarrow Y\}$ and $\{Y \Rightarrow X\}$ can be extracted, which satisfies $minsup$ and $minconf$; i.e., the association rule generation process can be deduced to a two-step operation; first, frequent pattern mining, and second, generating strong association rules of interest[18].

4.3 Incremental Approach to Data Mining for Frequent Pattern Extraction using FP-growth: Discovering Inter-Appliance Associations

In this subsection, the proposed approach towards progressive incremental frequent pattern mining along with additional interestingness measures for the discovery of correlation is discussed.

The mining of frequent patterns is generally considered as an off-line and costly process on large databases. In a real world application transaction data generation is a

continuous process, where new transactions are generated and old transactions may become obsolete as time progresses, thereby, invalidating existing frequent patterns and/or establishing new frequent pattern associations. Therefore, an incremental and progressive update strategy is imperative, where these variations/updates are taken into account and the discovered frequent patterns are duly maintained. For example, an appliance such as room-heater generally will be used during winter and we can expect reduced usage frequency during other seasons. As an effect, a significant gain during winter but decrease during other seasons will be registered. As a result, room-heater should appear higher on the list of frequent patterns and association rules during winter, but much lower during summer or spring. This objective can be achieved through progressive incremental data mining while eliminating the need to re-mine the entire database at regular intervals. Frequent pattern mining in a large database can be accomplished through pattern growth approach [79], [80]. We extend this pattern growth approach and present an incremental frequent pattern mining strategy of a progressive manner, which is discussed next.

4.3.1 FP-growth : A Pattern-Growth Approach, Without Candidate Generation For Mining Frequent Itemsets

Apriori [82] algorithm with candidate generation approach can suffer from the following problems:

- Breadth-first approach; i.e., level-wise search.
- Generate a large number of candidate sets.
- Repeatedly search through entire database to find support for an itemset.

To overcome these deficiencies, the work in [79] and [80] proposed pattern growth or FP-growth approach, which exploits depth-first divide-and-conquer technique. To start with, it generates a compact representation of the transactions from the database in the form of the frequent pattern tree or FP-tree. FP-tree preserves the association information, derived from each individual transaction, along with support count for each constituent item. Next, $conditional\ databases(tree)$ for each frequent item is extracted from the FP-Tree to mine frequent patterns, which the item under consideration is part of. This way, we are inspecting only the divided portion relevant to the item and its associated growing patterns, and addressing the shortcomings of the Apriori approach.

4.3.2 Incremental Frequent Pattern Extraction

Our proposed technique exploits the benefits of pattern growth strategy and extends it to achieve incremental progressive mining of frequent patterns by mining in a quantum of 24 hours; i.e., frequent patterns are extracted from data comprising of appliance usage tuples for a 24 hour period, in a progressive manner. With this approach, we mine only a portion of the entire database at each iteration, thus reducing the memory overhead for FP-growth strategy and achieve improved efficiency.

In our proposed approach, available data is recursively mined in quanta of 24 hours, and a frequent patterns discovered database, represented in table (4.2), is maintained across successive mining exercises. In other words, data mining can be viewed as a process conducted at the end of each day in an incremental manner. During each consecutive mining operation, the support count and database size for the existing frequent patterns are incremented and new patterns, with applicable support count and database

size, are added to persistent database. Moreover, we cease the use of minimum support threshold $minsup$ at the mining stage to eliminate any candidate patterns, resulting in discovery of all the possible frequent patterns. This change in technique is incorporated to avoid missing candidate patterns, which can become frequent if time quantum is increased or the complete database is mined in a single operation. At the end of the mining process, $database\_size$ is updated for all of the frequent patterns in the frequent patterns discovered database [table (4.2)] to ensure the correct computation of $support$.

Our proposed progressive incremental data mining approach is outlined by algorithm (1); which extends two-step process for frequent pattern mining using FP-growth. Step 1; i.e., FP-Tree construction is described in algorithm (3) and figures (4.1) and (4.2). Step 2; i.e., frequent pattern generation is covered in algorithm (5) and figures (4.3) and (4.4). Algorithm (2) explains the mechanism to achieve persistent storage of frequent patterns discovered by the mining process into a permanent storage such as a Data Base Management System.

Two step process for frequent pattern mining based on FP-growth

Step-1 Constructing Frequent Pattern Tree (FP-Tree): It takes two scans of transaction database; first to create the list of 1-itemset (an itemset comprising only one item) frequent itemsets with support [presented in table (4.1), sorted in decreasing order of support], and second to construct the FP-Tree[79], [80]. In our setting, we do not eliminate 1-itemset frequent itemsets based on minimum support threshold ($minsup$) for the reasons discussed earlier, which is a departure from the original algorithm proposed by [79], [80].

Algorithm 1 Incremental Frequent Pattern Mining

Require: Transaction database DB

Ensure: Incremental discovery of frequent patterns, stored in frequent patterns discovered database FP_DB

1: **for all** 24 hour quantum transaction data db_{24} in DB **do**
2: Determine database size $Database\_Size_{db_{24}}$ for quantum db_{24}
3: Constructing Frequent Pattern Tree (FP-Tree) {As described in Step-1}
4: Generating Frequent Patterns, while calling function $save\_update\_frequent\_pattern$ to save frequent patterns to frequent patterns discovered database FP_DB {As described in Step-2}
5: **end for**
6: For all Frequent Patterns in Database FP_DB increment $Database\ Size$ by $Database\_Size_{db_{24}}$

Algorithm 2 Function $save\_update\_frequent\_pattern$

Require: Frequent Pattern extracted $FP\_extracted$, support count $absolute\_support$, Frequent pattern discovered database FP_DB

Ensure: Add or update Frequent Pattern in frequent patterns discovered database

1: Search a frequent pattern $FP = FP\_extracted$ in FP_DB
2: **if** Frequent Pattern found **then**
3: Increment support count by $absolute\_support$.
4: **else**
5: Add a new Frequent Pattern with support count $absolute\_support$ and $Database\_size$ = 0.
6: **end if**

TABLE 4.1: List: 1-itemset frequent itemsets with support

1-itemset	12	2	4	3
support	8	7	6	5

2 = Laptop, 3 = Monitor
4 = Speakers
12 = Washing Machine

Decreasing order of support

FP-Tree is constructed by reading one transaction at a time from Frequent Pattern Source Database [table (3.1)], sorting the items according to the list of 1-itemsets [table (4.1)] in decreasing order of global support, and mapping it to a path in the FP-Tree. Fixed order of items ensures an overlap of path for the transactions having identical prefix; i.e., sharing items. Further, on the addition of a new transaction, the support count is incremented for each item node in the prefix path shared among transactions. A header table comprising of 1-itemsets, sorted in decreasing order of global support, is maintained. This table stores pointers to the last item added in the tree of particular 1-itemset. Additionally, a $node\text{-}link$ is added from recently added item node to the preceding item node of the same 1-itemset type. $node\text{-}link$ facilitates traversal (trace) of all the 1-itemsets in FP-Tree for a specific item. Therefore, an FP-Tree is a compact representation of the database which preserves the critical information of $absolute\ support$ for each item/itemset and transaction patterns. This process is depicted in figure (4.3) by addition of two transactions to FP-Tree. Figure (4.4) presents the final FP-Tree constructed from Frequent Pattern Source Database [table (3.1)].

Algorithm 3 Step-1 Constructing FP-Tree

Require: Given transaction database DB

Ensure: FP-Tree T

1: Scan DB, generate list F with all the 1-itemset frequent items, and determine support of each frequent item. {Database Scan 1: create list of 1-itemset frequent itemsets with support}
2: Sort F in descending order of support.
3: Create FP-Tree root $T \leftarrow null$.
4: **for all** Transactions $(Tr) \in DB$ **do** {Database Scan 2: create FP-Tree}
5: Sort items in Tr according to order of F.
6: Item list of pattern $[P = p|P']$, where p is first element and P' remaining list
7: Call Function $insert\_tree([P = p|P'], T)$.
8: **end for**

Algorithm 4 Function $insert\_tree([p|P'], T)$

Require: frequent item list of pattern $[P = p|P']$, where p is first element and P' remaining list, and FP-Tree T.

Ensure: Add items from an item list to FP-Tree T.

1: Initialize current root node $R_C \leftarrow root\_node$ (FP-Tree).
2: **for all** Item/element $item_i \in P$ **do**
3: Search a node N in T, having N = $item_i$.
4: **if** Node found **then**
5: Increment support count for N by 1.
6: $R_C \leftarrow N$ {Capture as current root}
7: **else**
8: Create new node N, with support count 1; where parent-link linked to R_C, and node-link linked to nodes with same item.
9: $R_C \leftarrow N$ {Capture as current root}
10: **end if**
11: **end for**

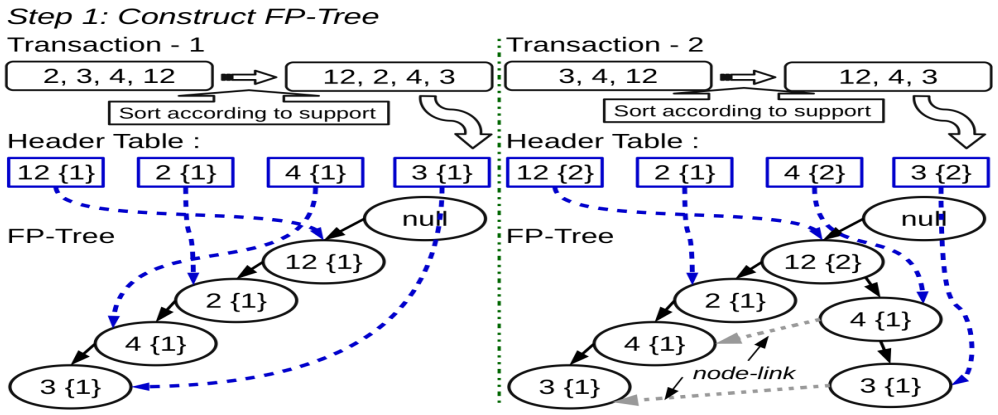


FIGURE 4.1: Step 1 - FP-Tree Construction : Scanning database and adding transactions

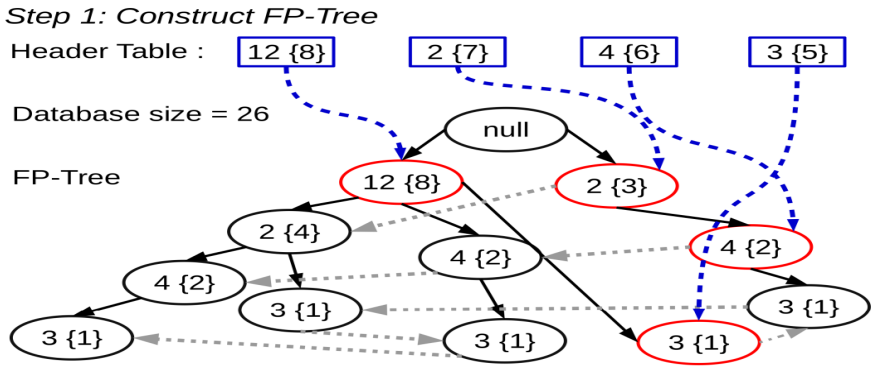


FIGURE 4.2: Step 1 - FP-Tree Construction : Final frequent pattern tree

Step-2 Generating Frequent Patterns: Once FP-Tree is created, a bottom-up recursive elimination approach making use of the divide and conquer scheme is taken to generate a complete set of frequent patterns from the FP-Tree. the FP-tree mining is accomplished by starting from each frequent length-1 pattern (as an initial suffix pattern), and constructing it's *conditional-pattern-base* from prefix paths extracted from the FP-tree, where suffix co-occur with prefix. *conditional-pattern-base* can be considered as a collection of transactions containing a particular itemset, but removing it from transactions. The header

table, as shown in figure (4.2) acts as the source for 1-itemsets; which are ordered in the increasing order of global support, so the process starts from the leaf (bottom) nodes of the tree and traverses towards the root (up). Later, $conditional\text{-}FP\text{-}tree$ is created from the $conditional\text{-}pattern\text{-}base$ and mined recursively to extract frequent patterns until the resulting tree is empty or comprises of a single path. Lastly, frequent pattern from single paths are derived by producing all of the combinations of the sub-paths. The pattern growth is accomplished through concatenation of the suffix pattern with the frequent patterns generated from a conditional FP-tree. $node\text{-}link$ enables extraction of $conditional\text{-}pattern\text{-}base$ and creation of $conditional\text{-}FP\text{-}tree$ by aiding the trace of nodes for a given 1-itemset suffix pattern, which is frequent. Figures (4.3) and (4.4) illustrate the procedure for the extraction of frequent patterns.

The results of frequent pattern mining are represented in table For the purpose of making frequent patterns available for future manipulation and utilization, we store all the frequent patterns extracted into a Database Management System.(4.2). Thus, a frequent pattern of this form, for example "Laptop, Monitor, Speakers", can be interpreted to represent association among appliances or inter-appliance associations. The respective probabilistic evidence, supporting occurrence of these association, can be computed as shown in equations (4.1) and (4.2).

TABLE 4.2: Frequent Patterns: frequent patterns discovered database

Frequent Pattern	Absolute Support	Database Size
'2 3'	3939	7899
'2 4'	2840	7899
'2 3 4'	2649	7899
'3 4'	2649	7899
'2 3 4 12'	1299	7899
2 = Laptop, 3 = Monitor, 4 = Speakers, 12= Washing Machine		

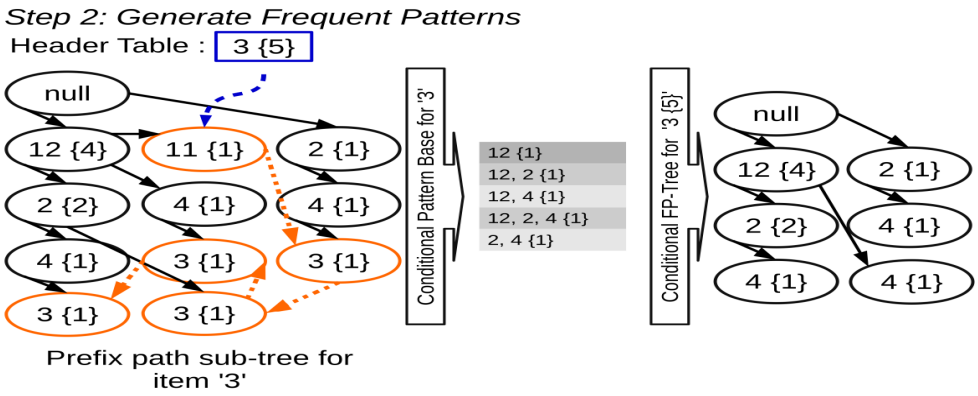


FIGURE 4.3: Step 2 - Generating Frequent Pattern : Conditional FP-Tree

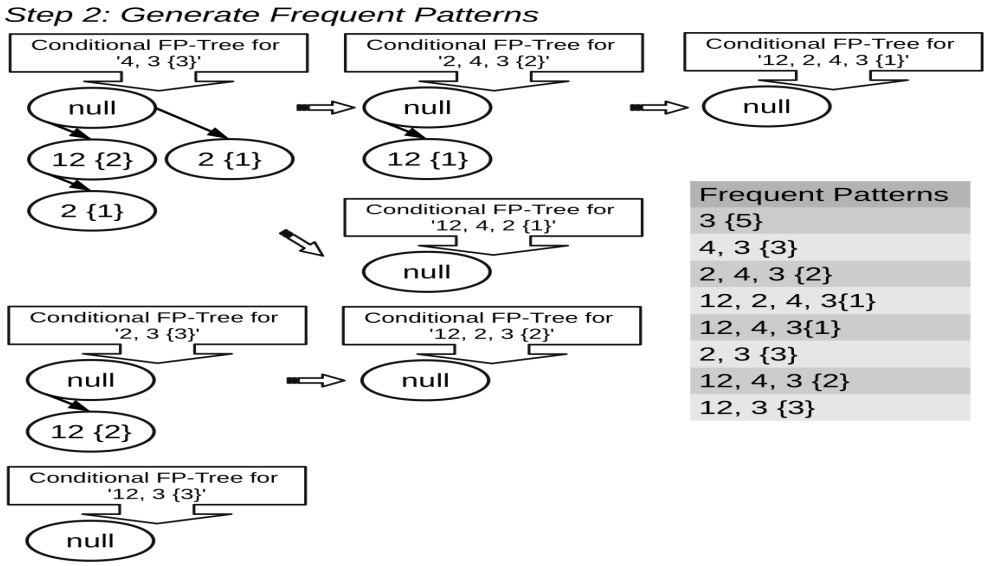


FIGURE 4.4: Step 2 - Generating Frequent Pattern : Recursive mining

Algorithm 5 Step-2: FP-growth: Generating Frequent Patterns.

Require: FP-Tree T, Current itemset suffix S.

Ensure: Frequent Patterns

1: **if** T is a single path **then** {Mine single path FP-Tree for frequent patterns}
2: **for all** Combination C of nodes in T **do** {support(C) = minimum support of nodes C}
3: Generate frequent patterns $FP_S = C \cup S$
4: update Frequent Pattern Discovered Database FP_DB
5: $FP\_single\_path = FP\_single\_path \cup FP_S$
6: **end for**
7: FP_single_path represents Frequent patterns generated
8: **else**
9: T is multipath tree {Mine multipath FP-Tree for frequent patterns}
10: **for all** $item_i$ of nodes in T **do**
11: Generate frequent pattern $FP_M = item_i \cup S$
12: update Frequent Pattern Discovered Database FP_DB
13: $FP\_multipath = FP\_multipath \cup FP_M$
14: Determine itemset suffix $S_i = item_i \cup S$
15: Extract conditional prefix path or conditional pattern-base for $item_i$ by using node-link and parent-link.
16: Generate conditional FP-Tree T_i from conditional prefix path or conditional pattern-base.
17: **if** FP-Tree $T_i \neq \emptyset$ **then**
18: Call FP-growth(FP-Tree T_i, Current itemset suffix S_i)
19: **end if**
20: **end for**
21: $FP\_multipath$ represents Frequent patterns generated
22: **end if**
23: Final set of frequent patterns generated = $FP\_single\_path \cup FP\_multipath$

4.3.3 Association Rules Generation Using Correlation Analysis

In a large database the number of association rules generated can be very large. Although, it is entirely dependent on the application of results, but reduction in the number of association rules can help narrow down the search space for the useful and strong association rules. This can be achieved through application of statistical interestingness measures. In general, FP-growth [79], [80] and Apriori [82] algorithms use $support$-$confidence$ frame-work to generate frequent patterns and extract association rules, while eliminating uninteresting rules by comparing $support$ and $confidence$ with $minsup$ and $minconf$ respectively. However, $support$ and $confidence$ do not evaluate correlation of the rule's $antecedent$ and $consequent$. This turn out to be less effective in eliminating uninteresting association rules. Therefore, it is important to learning correlation relationship among the rules constituents to determine positive or negative impact of one's presence over other and remove the rules which are not of interest. A measure of correlation such as $Lift$, Kulczynski measure ($Kulc$) and/or imbalance ratio (IR) can help and supplement the $support$-$confidence$ frame-work and provide more insight into the association relationship [18]. Consequently, the correlation rule can be expressed as defined in equation (4.4).

Lift: measures dependency and correlation of rule's $antecedent$ and $consequent$; $lift$ is define as,

$$X \Rightarrow Y [support, confidence, correlation] \tag{4.4}$$

$$\begin{aligned} lift(X,Y) &= \frac{P(X \cup Y)}{P(X).P(Y)} \\ &= \frac{P(Y|X)}{P(Y)} \\ &= \frac{confidence(X \Rightarrow Y)}{support(Y)} \end{aligned} \tag{4.5}$$

If $Lift < 1.0$, then X is negatively correlated to Y; i.e., occurrence of X indicates absence of Y or vice-versa. If $Lift > 1.0$, then X and Y are positively correlated; i.e., occurrence of X indicates presence of Y or vice-versa; where-as if $Lift = 1.0$, then X and Y are independent with no correlation among them.

This commonly used correlation measure $Lift$ is affected by null-transactions. Null-transaction, are transactions where itemsets under consideration are not part of it ($X \in$ null-transaction *or* $Y \in$ null-transaction), and in a large database null-transactions can outbalance the support count for itemsets. Hence, this approach fails when contemplating low minimum support threshold or searching for extended patterns as explained by the study in [18]. This study suggests using null-invariant interestingness measures of Kulczynski measure ($Kulc$) along with Imbalance Ratio (IR) to supplement $support$-$confidence/lift$ frame-work to extract more interesting rules.

Kulczynski Measure ($Kulc$)[18]: $Kulc$ of X and Y, is an average of $confidence$ measures for X and Y; which, by definition of $confidence$ can be translated into average of conditional probabilities. $Kulc$ measure is null-invariant and is defined as:

$$Kulc(X,Y) = \frac{1}{2}(P(X|Y) + P(Y|X)) \tag{4.6}$$

If $Kulc = 0.0$, then X is negatively correlated Y; i.e., occurrence of X indicates absence Y or vice-versa. If the $Kulc = 1.0$, then X and Y are positively correlated; i.e., occurrence X indicates presence Y or vice-versa, whereas $Kulc = 0.50$ indicates X and Y are independent having no correlation.

Imbalance Ratio (IR)[18]: IR measures the imbalance of $antecedent$ and $consequent$ for the rule. It is defined as,

$$IR(X,Y) = \frac{|s_X - s_Y|}{s_X + s_Y - s_{(X \cup Y)}} \tag{4.7}$$

Where $IR = 0.0$ and $IR = 1.0$ represent perfectly balanced and very skewed scenario respectively. Imbalance ratio is null-invariant and is not influenced by the database size.

Algorithm 6 Apriori Association Rule Generation

Require: Frequent patterns discovered database FP_DB, minimum support $minsup$, minimum confidence $minconf$, minimum Kulczynski Measure $minkulc$

Ensure: Association Rules

1: **for all** Frequent itemset FP_i in FP_DB **do**
2: Generate all subsets $subset_{FP_i}$ having $subset_{FP_i} \neq \emptyset$
3: **for all** Subset in $subset_{FP_i}$ **do**
4: $X \leftarrow subset_{FP_i}$
5: $Y \leftarrow (FP_i - subset_{FP_i})$
6: $support(s_{X \Rightarrow Y}) \leftarrow support(X \cup Y)$
7: $confidence(c_{X \Rightarrow Y}) \leftarrow \frac{support(X \cup Y)}{support(X)}$
8: $Kulc \leftarrow \frac{1}{2}\{confidence(X \Rightarrow Y) + confidence(Y \Rightarrow X)\}$
9: **if** $support(s_{X \Rightarrow Y}) \geq minsup$ **AND** $confidence(c_{X \Rightarrow Y}) \geq minconf$ **AND** $Kulc \geq minkulc$ **then**
10: Output rule "$subset_{FP_i} \Rightarrow (FP_i - subset_{FP_i})$"
11: **end if**
12: **end for**
13: **end for**

Association Rule Generation: It is an effortless process to extract association rules from frequent itemsets discovered from transactions in a database DB. Association rules can be derived, as explained in algorithm (6), where we introduce the use of the correlation measures of $Kulc$ to extend the Apriori [82] approach in order to filter out uninteresting association rules along with the measure of imbalance ratio IR to explain it.

4.4 Results

4.4.1 Results - Summary

Exhaustive incremental frequent pattern mining was carried out using the energy consumption data from five houses in the dataset UK-Dale [8] and one house in the dataset AMPds2 [9], and the synthetic dataset to examine intermediate and final results. Moreover, a comprehensive analysis of power load and energy consumption patterns was conducted to verify and explain these results. The results from three houses representative of the findings are presented. Table (4.3) summarizes the various resulting outcomes and components with respective interpretation. A detailed discussion is presented in next section (4.4.2).

TABLE 4.3: Summary of Results

Figure/Tables	Description
Figures: 4.5, 4.6, 4.7, 4.8, 4.9, 4.10, 4.11, 4.12, 4.13, 4.14	Incremental frequent pattern mining intermediate and final results, represented in a tree structure visualizing associative relationship among appliances with respective $support$
Table: 4.4	Incremental progressive inter-appliance association discovery
Table: 4.5	Discovered appliance association/correction rules .
Table: 4.6	Appliance usage priority learned for the premises
Figures: 4.15a, 4.15b, 4.16a	• Energy Consumption Analysis - Average power load, average energy consumption and aggregate energy consumption analysis for the premises • Average power vs. Energy consumption vs. Usage duration for appliances contributing to peak load and peak usage (hours) • Appliance energy consumption contribution at peak load hours • Appliance energy consumption contribution at peak energy consumption hours
Figure 4.17	Representative energy consumption curve for House 2, including four appliances that is Washing Machine, Laptop, Monitor and Speakers.

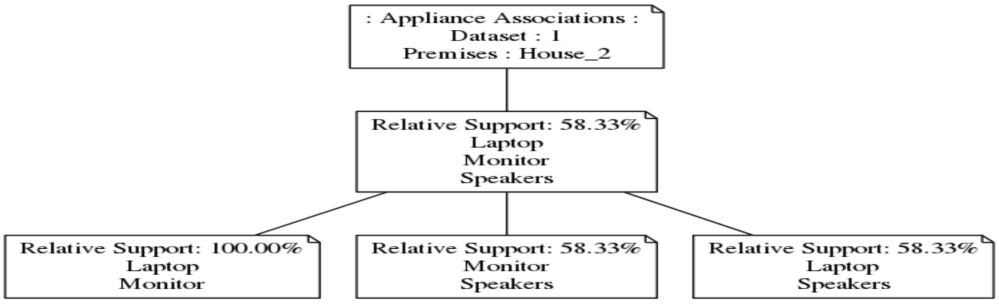


FIGURE 4.5: House 2:Appliance Associations upto 1 days data ($minsup \geq 0.2$)

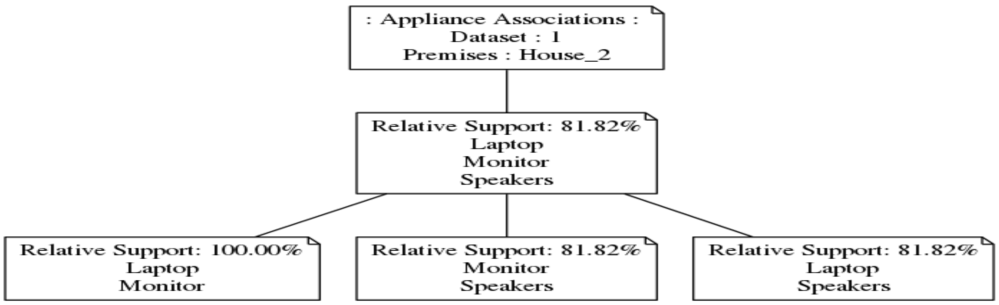


FIGURE 4.6: House 2:Appliance Associations upto 2 days data ($minsup \geq 0.2$)

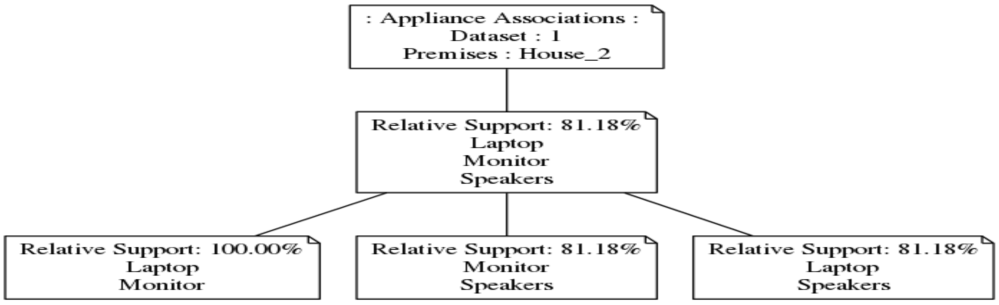


FIGURE 4.7: House 2:Appliance Associations upto 3 days data ($minsup \geq 0.2$)

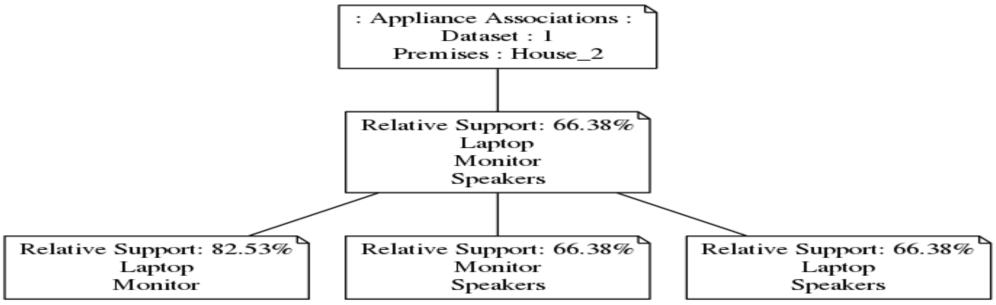


FIGURE 4.8: House 2:Appliance Associations upto 7 days data ($minsup \geq 0.2$)

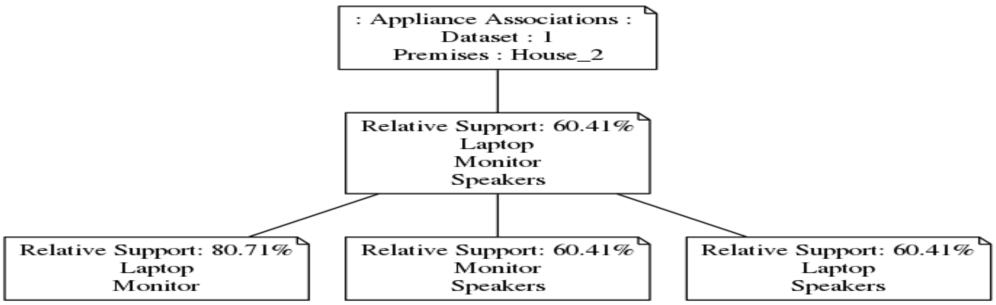


FIGURE 4.9: House 2:Appliance Associations upto 15 days data ($minsup \geq 0.2$)

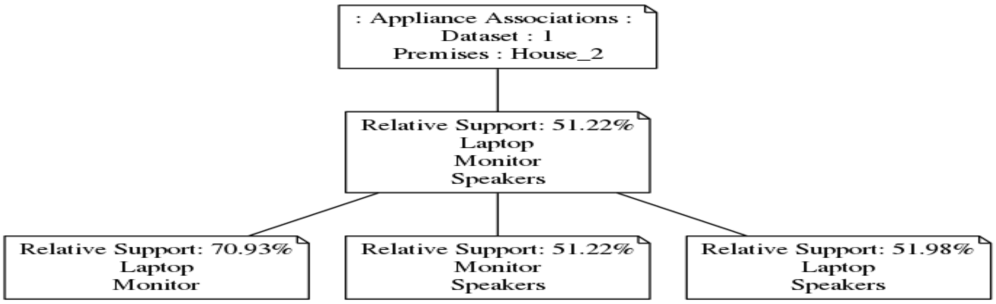


FIGURE 4.10: House 2:Appliance Associations upto 30 days data ($minsup \geq 0.2$)

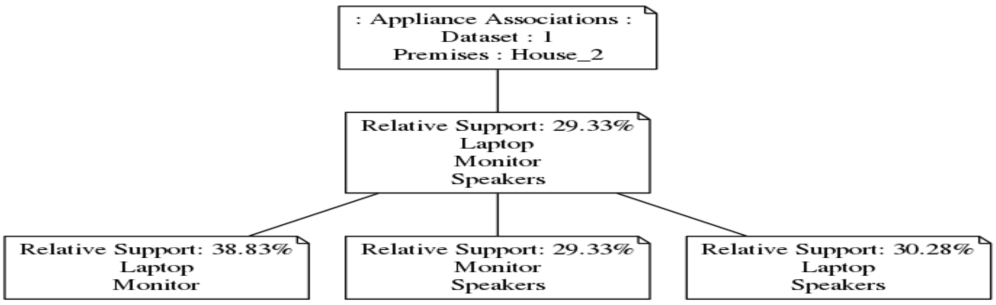


FIGURE 4.11: House 2:Appliance Associations upto 25% dataset ($minsup \geq 0.2$)

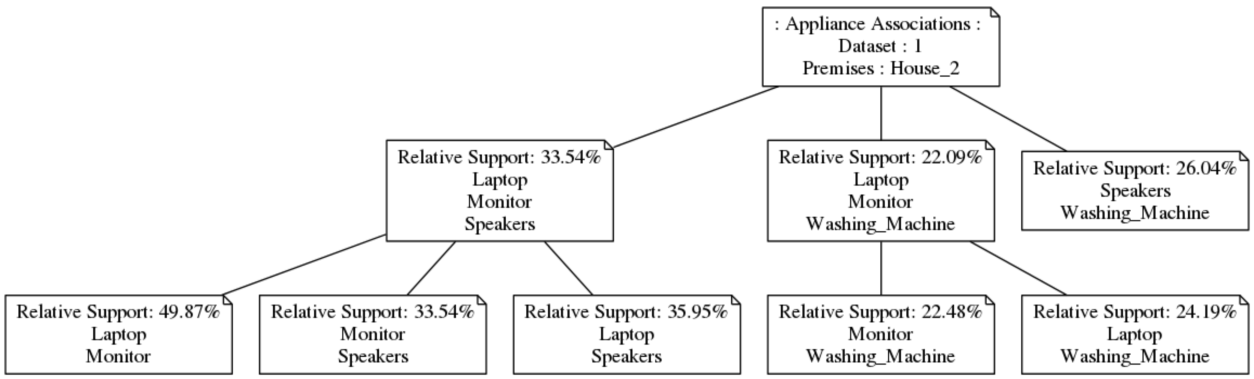


FIGURE 4.12: House 2:Appliance Associations in full database ($minsup \geq 0.2$)

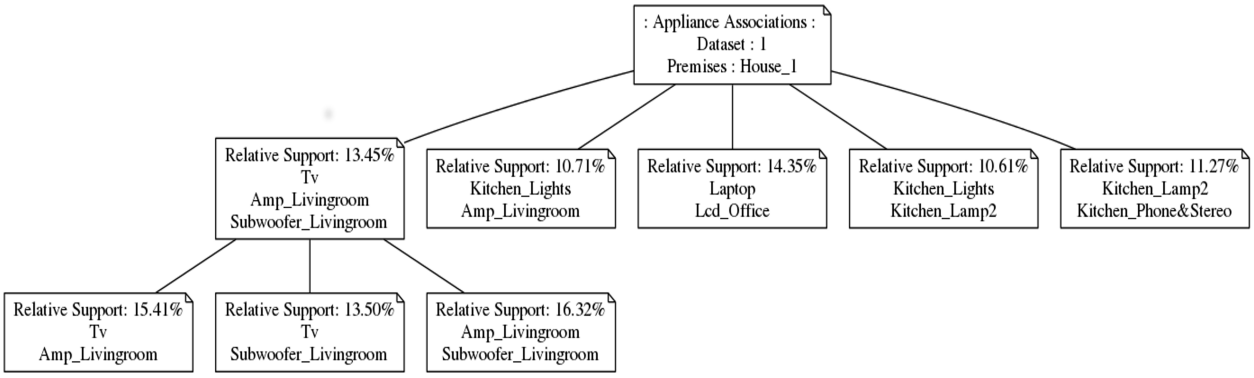


FIGURE 4.13: House 1:Appliance Associations in full database ($minsup \geq 0.1$)

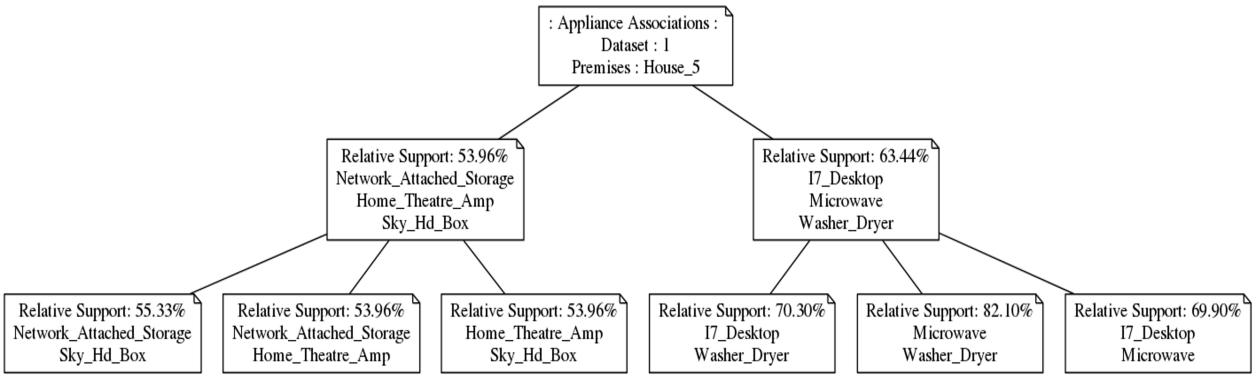


FIGURE 4.14: House 5:Appliance Associations in full database ($minsup \geq 0.5$)

TABLE 4.4: Appliances Associations

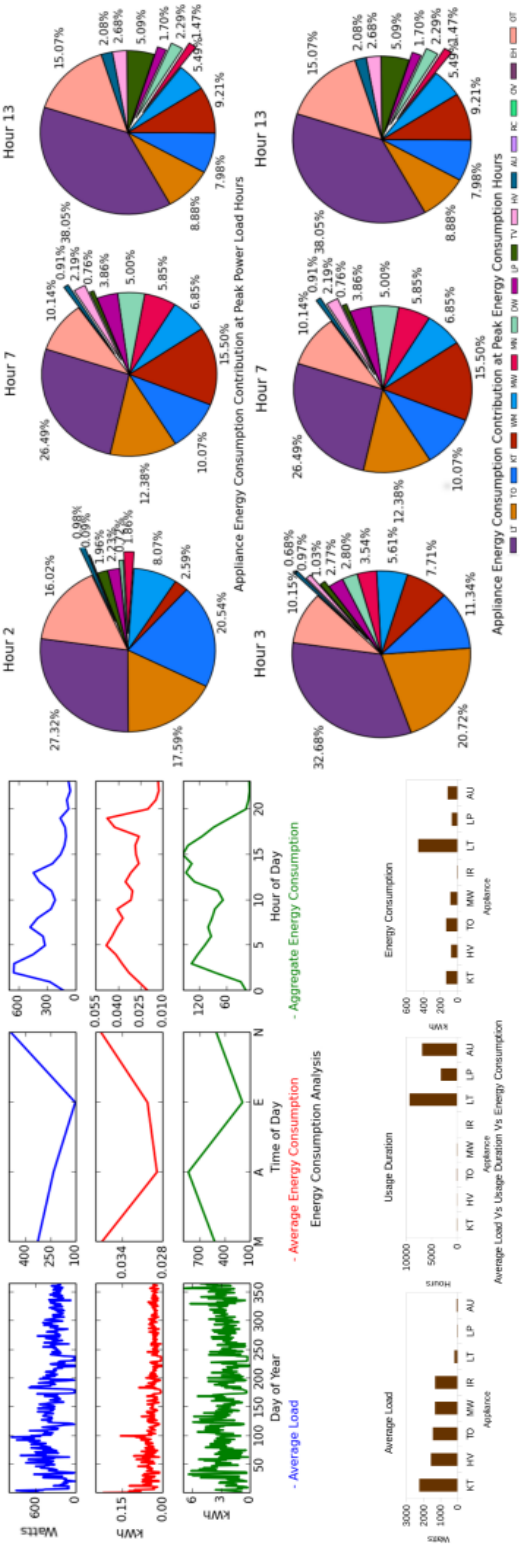
Incremental Progressive Inter-Appliance Association Discovery

	Appliances ↓	Support (%)			
	Database size/Time Period* →	7 days	30 days	25 %	100 %
House 1	Washing_Machine, Kitchen_Lights	8.11			
	Kitchen_Lights, Kettle	8.11			
	Laptop, Kitchen_Lights		24.32	7.11	9.10
	Laptop, Lcd_Office				14.35
	Kitchen_Lights, Livingroom_S_Lamp				7.36
	Kitchen_Lights, Kitchen_Lamp2				10.61
	Livingroom_S_Lamp, Kitchen_Lamp2				7.43
	Kitchen_Lamp2, Kitchen_Phone&Stereo				11.27
	Lcd_Office, Office_Lamp3				7.00
	TV, Amp_Livingroom, Subwoofer_Livingroom, Livingroom_Lamp_TV				7.00
	TV, Subwoofer_Livingroom, Livingroom_Lamp_TV				7.03
	TV, Subwoofer_Livingroom				13.50
	TV, Livingroom_Lamp_TV				7.12
	Subwoofer_Livingroom, Livingroom_Lamp_TV				7.85
	TV, Amp_Livingroom, Subwoofer_Livingroom				13.45
	TV, Amp_Livingroom			7.05	15.41
	Amp_Livingroom, Subwoofer_Livingroom				16.32
	TV, Amp_Livingroom, Livingroom_Lamp_TV				7.09
	Amp_Livingroom, Livingroom_Lamp_TV				7.77
	Amp_Livingroom, Subwoofer_Livingroom, Livingroom_Lamp_TV				7.63
	TV, Kitchen_Lights, Amp_Livingroom				7.42
	TV, Kitchen_Lights				7.70
	Kitchen_Lights, Amp_Livingroom				10.71
	Kitchen_Lights, Amp_Livingroom, Subwoofer_Livingroom				8.18
	Kitchen_Lights, Subwoofer_Livingroom				9.59
House 2	Laptop, Monitor, Speakers	66.38	51.22	29.33	33.54
	Laptop, Monitor	82.53	70.93	38.83	49.87
	Monitor, Speakers	66.38	51.22	29.33	33.54
	Laptop, Speakers	66.38	51.98	30.28	35.95
	Laptop, Monitor, Washing Machine				22.09
	Monitor, Washing Machine				22.48
	Laptop, Washing Machine				24.19
	Speakers, Washing Machine				26.04
House 5**	Network_Attached_Storage, Home_Theatre_Amp, Sky_HD_box				53.96
	Network_Attached_Storage, Sky_HD_box				55.33
	Network_Attached_Storage, Home_Theatre_Amp				53.96
	Home_Theatre_Amp, Sky_HD_box				53.96
	I7_Desktop, Microwave, Washer_Dryer				63.44
	I7_Desktop, Washer_Dryer			68.71	70.30
	Microwave, Washer_Dryer				82.10
	I7_Desktop, Microwave			63.96	69.90

House 1: $minsup \geq 0.07$, House 2: $minsup \geq 0.2$, House 5: $minsup \geq 0.5$

* - Database size/Time Period - the training data size, which is processed

** - Large result-set for 7 and 30 days; hence excluded.

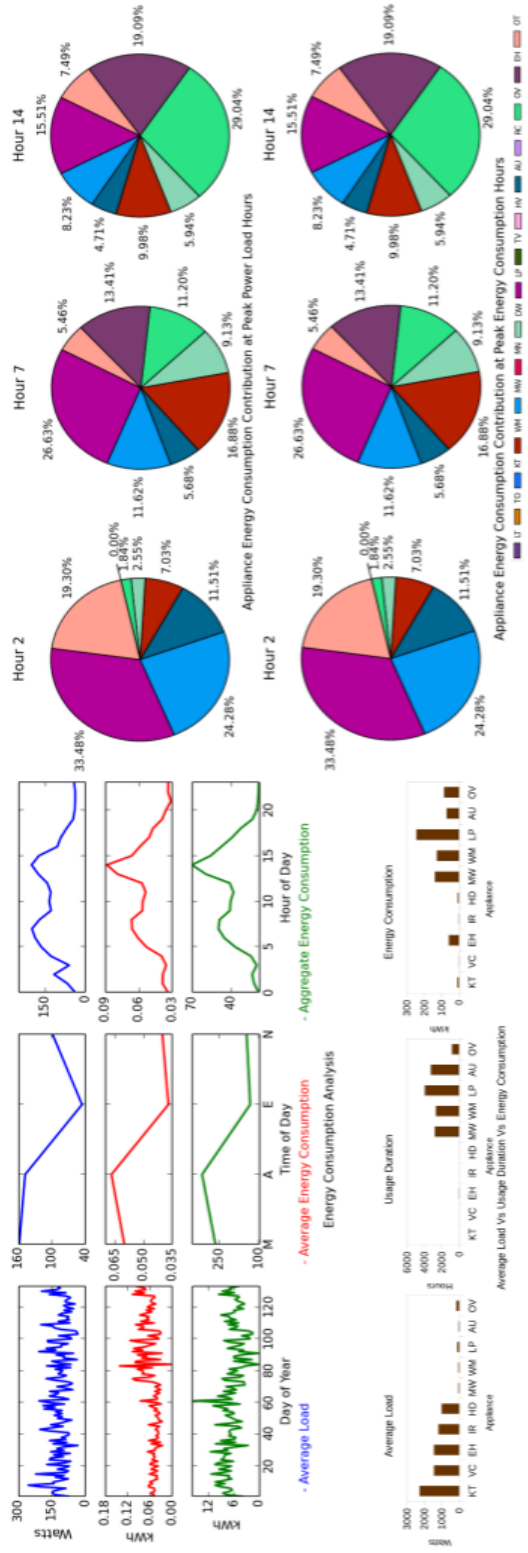


(A) House 1 : Energy consumption analysis

(B) House 2 : Energy consumption analysis

FIGURE 4.15: Energy consumption analysis : House 1 and House 2

[KT - Kettle, HV - Hoover, TO - Toaster, MW - Microwave, IR - Iron, LT - Lights, LP - Laptop or Desktop, AU - Audio/music equipment : Home theater or Amp or Speakers, MN - Monitor, VC - Vacuum Cleaner, EH - Electric Hob, HD - Hairdryer, WM - Machine Machine or Dryer, OV - Oven]



(A) House 5 : Energy consumption analysis

FIGURE 4.16: Energy consumption analysis : House 5

[KT - Kettle, HV - Hoover, TO - Toaster, MW - Microwave, IR - Iron, LT - Lights, LP - Laptop or Desktop, AU - Audio/music equipment : Home theater or Amp or Speakers, MN - Monitor, VC - Vacuum Cleaner, EH - Electric Hob, HD - Hairdryer, WM - Machine Machine or Dryer, OV - Oven]

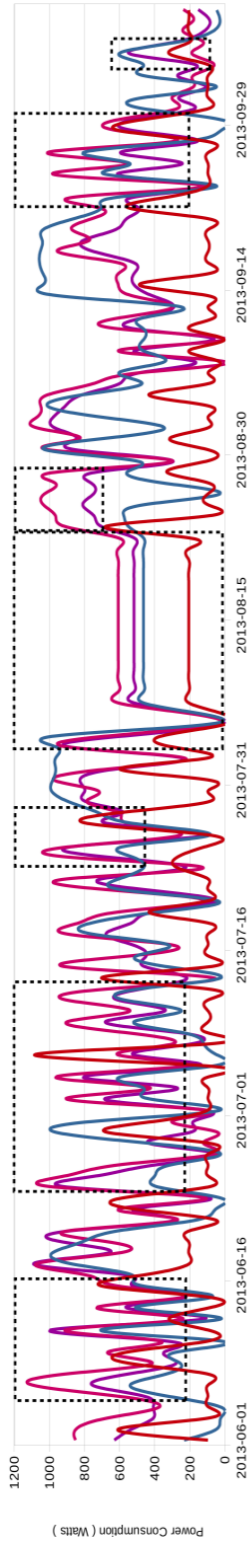


FIGURE 4.17: House 2: Appliances Association via Energy Curves

Appliances : Laptop, Monitor, Speakers, Washing Machine

TABLE 4.5: Appliance Association Rules

	Sr.	Association Rule	Support	Confidence	Kulc	IR
House 1	1	Tv ⇒ Amp_Livingroom	0.15	0.96	0.85	0.23
	2	Amp_Livingroom ⇒ Tv	0.15	0.73	0.85	0.23
	3	Lcd_Office ⇒ Laptop	0.14	0.93	0.75	0.38
	4	Tv, Amp_Livingroom ⇒ Subwoofer_Livingroom	0.13	0.87	0.78	0.19
	5	Kitchen_Lights, Subwoofer_Livingroom ⇒ Amp_Livingroom	0.08	0.85	0.62	0.51
	6	Amp_Livingroom, Livingroom_Lamp_Tv ⇒ Subwoofer_Livingroom	0.08	0.98	0.69	0.6
	7	Subwoofer_Livingroom, Livingroom_Lamp_Tv ⇒ Amp_Livingroom	0.08	0.97	0.67	0.62
House 2	1	Monitor ⇒ Laptop	0.50	0.99	0.96	0.06
	2	Laptop ⇒ Monitor	0.50	0.93	0.96	0.06
	3	Speakers ⇒ Laptop	0.36	0.74	0.70	0.08
	4	Monitor, Speakers ⇒ Laptop	0.34	1.00	0.81	0.38
	5	Laptop, Speakers ⇒ Monitor	0.34	0.93	0.80	0.27
	6	Monitor, Washing Machine ⇒ Laptop	0.22	0.98	0.70	0.58
	7	Laptop, Washing Machine ⇒ Monitor	0.22	0.91	0.68	0.50
	8	Laptop, Washing Machine ⇒ Speakers	0.18	0.74	0.56	0.45
	9	Monitor, Speakers, Washing Machine ⇒ Laptop	0.16	1.00	0.65	0.69
	10	Laptop, Speakers, Washing Machine ⇒ Monitor	0.16	0.91	0.62	0.62
	11	Monitor, Washing Machine ⇒ Laptop, Speakers	0.16	0.73	0.59	0.32
House 5	1	Washer_Dryer ⇒ Microwave	0.82	0.92	0.9	0.04
	2	Microwave ⇒ Washer_Dryer	0.82	0.88	0.9	0.04
	3	I7_Desktop ⇒ Washer_Dryer	0.70	0.91	0.85	0.13
	4	Washer_Dryer ⇒ I7_Desktop	0.70	0.79	0.85	0.13
	5	I7_Desktop ⇒ Microwave	0.70	0.91	0.83	0.16
	6	Microwave ⇒ I7_Desktop	0.70	0.75	0.83	0.16
	7	I7_Desktop Microwave ⇒ Washer_Dryer	0.63	0.91	0.81	0.2
	8	I7_Desktop Washer_Dryer ⇒ Microwave	0.63	0.90	0.79	0.23
	9	I7_Desktop ⇒ Microwave Washer_Dryer	0.63	0.83	0.8	0.05
	10	Microwave Washer_Dryer ⇒ I7_Desktop	0.63	0.77	0.8	0.05
	11	Washer_Dryer ⇒ I7_Desktop Microwave	0.63	0.71	0.81	0.2

House 1 : $Kulc \geq 0.60$, $minsup \geq 0.08$, $minconf \geq 0.70$, House 2 : $Kulc \geq 0.55$, $minsup \geq 0.10$, $minconf \geq 0.70$
House 5 : $Kulc \geq 0.75$, $minsup \geq 0.60$, $minconf \geq 0.70$

4.4.2 Results - Analysis and Discussion

Table (4.4), shows the incremental appliance-appliance associations discovery; it was noted that the associative relationships change over a period of time and revealed changing personal preferences that were indicative of consumer comfort in relation to the use of electrical appliances. Formation of appliance-appliance associations were duly revealed by intermediate results, which strengthens during each step of incremental mining. Figures (4.5-4.12), (4.13), and (4.14) illustrated the visualization of progressive and incremental development of appliance associations using a tree structure show the associative relationship among appliances with respective $support$ in the database. Tables (4.5) and (4.6)

TABLE 4.6: Appliance Usage Priority

	Sr.	Appliance	Relative Support (%)
House 1	1	Kitchen_Lights	42.82
	2	Kitchen_Lamp2	31.10
	3	Laptop	25.56
	4	Amp_Livingroom	21.01
	5	Subwoofer_Livingroom	19.46
	6	Livingroom_S_Lamp	18.08
	7	Kitchen_Phone&Stereo	17.86
	8	LCD_Office	15.44
	9	Livingroom_Lamp_TV	11.08
	10	Office_Lamp2	9.28
	11	Washing_Machine	8.63
	12	Livingroom_S_Lamp2	8.16
	13	Office_Lamp3	7.72
	14	Kettle	5.77
House 2	1	Washing Machine	62.74
	2	Laptop	53.84
	3	Monitor	50.34
	4	Speakers	48.83
	5	Laptop 2	15.43
	6	Running Machine	09.61
	7	Kettle	06.13
House 5	1	Microwave	92.86
	2	Washer_Dryer	88.96
	3	I7_Desktop	76.89
	4	Network_Attached_Storage	55.33
	5	Sky_Hd_Box	55.33
	6	Home_Theatre_Amp	53.96

House 1 : $minsup \geq 0.05$, House 2 : $minsup \geq 0.05$
House 5 : $minsup \geq 0.50$

present the inter-appliance association rules and appliance usage priorities, which constitute behavioral appliance usage patterns establishing household energy usage patterns.

For House 2, four appliances (Washing Machine, Laptop, Monitor and Speakers) exhibited strong association rules, and the energy consumption curve complimented the results of the frequent pattern and association rules discovery. It was noted that the representative energy consumption patterns of different appliances, refer marked portion in energy consumption curves figure (4.17), exhibited close similarities, which was expected. Thus, it can be inferred that these appliances were used simultaneously as shown by the data mining outcomes. Additionally, consumer behavioral traits were revealed by the

frequent patterns and association rules. For example, Laptop was used along with Washing Machine; i.e., household occupants often liked to work on computer while washing clothes and listening to music. Based on appliance usage priorities, the Washing Machine was found to be the most used appliance for the House 2.

Analogous results were obtained for House 1 and House 5. For House 1, Kitchen _Lights, Subwoofer, Amp, and TV shown to be strongly associated, which indicates that the occupants often watched TV or listened to music while cooking. The Kitchen_Lights were the most used appliance in the house. House 5 results showed that the Microwave, Washer_Dryer, and I7_Desktop display were strongly associated with the Microwave being the most frequently used appliance. it can be inferred that the occupants at House 5 enjoyed working on the computer while cooking or washing and drying cloths. Furthermore, it was observed that season, month, week, weekday, time of day and hour of day time frames had strong influences on occupant behavior and an immediate impact on energy usage, based on appliance usage patterns obtained through incremental data mining. Therefore, it is critical to learn these variations at regularly intervals, most suitably near-real time, to take these variations into account while designing energy programs.

Table (4.6) shows appliance usage priorities, derived from the frequency of use of respective appliance, which can determine the Appliance of Interest AoI. AoI's are appliances most frequently used and with higher usage periods, these are as major contributors to household energy consumption. AoIs have small power ratings but contribute large portion of energy consumption compared to appliances with higher power ratings. Appliances with high power ratings contribute to the peak power (kW) load but overall energy (kWh) consumption is higher for appliances identified as AoIs, because of short or very short operating time and long usage durations, respectively for these groups of

appliances.

Extensive energy consumption, peak load, peak energy usage, and appliance usage duration analyses were conducted to verify the results and determined homogeneous trends. Figures (4.15a), (4.15b), and (4.16a), illustrate the energy consumption analysis conducted for three houses with trend plots for the average load (Watts), average energy consumption (kWh) and aggregate energy consumption (kWh) on daily, time of day and hour of day scales. From the trend plots the peak load (Watts) and peak energy consumption (kWh) hours can be identified. The bar chart analyses show the average load (Watts) against usage duration (Hours) and energy consumption (kWh) for two group of appliances (i.e. appliances contributing to peak load (kW) and appliances with the highest usage duration for the house). Additionally, the pie charts illustrate the analysis of the appliances energy consumption contribution during peak hours using is peak load (kW) and peak energy consumption (kWh). Based on the energy consumption analysis, the appliances contributing to peak power loads were Kettle, Toaster, Microwave, and Rice_Cooker, although these were not the only appliances that contributed to household energy consumption. In contrast, the AoIs such as Monitor, Laptop, and Speakers, had small power ratings but contributed larger portions of the energy usage for each household. For example, in House 2 appliances Kettle, Toaster, and Microwave had high power rating but major portion of energy usage was contributed by Monitor and Laptop despite being small power footprint appliances. Similar observations were noted for House 1, where appliances Hoover, Toaster, and Microwave had high power ratings but consumed relatively small amounts of energy compared to Lights and Laptop, which were low power appliances. Additionally, for House 5, appliances Laptop, Audio/music equipment, and Oven made-up group of AoIs against high power rating appliances such as

Kettle, Vacuum Cleaner, Electric Hob, and Hairdryer.

Appliances such as the Washer and Dryer, which are the first candidates for various demand side management programs but were not the major contributors to household energy consumption; rather, manually operated appliances such as Kettle, Lights, and Laptop were the largest contributors to energy usage. This information can assist in establishing individual households' behavioral attributes or household energy profiles more effectively and accurately. The success of energy programs relies on increasing active participation among consumers, and these programs will require to conform more accurately to consumer personal preferences over time to achieve it. In addition, appliance-time association results, which are presented in Chapter (5), support these findings and represent the usage concentration of AoIs during peak load and peak energy consumption hours.

The outcome of the evaluation comprehensively supported the research hypothesis of the undeviating influence of human behavior on energy consumption patterns at a household level, which can be learned through inter-appliance associations.

Chapter 5: Progressive Incremental Data Mining : Cluster Analysis

5.1 Cluster Analysis - Incremental k-means to Discover Appliance-Time Associations

In addition to uncovering inter-appliance associations, understanding the appliance-time associations can aid critical analysis of consumer energy consumption behavior with respect to preferences on time of energy usage. Appliance-time associations can be defined with respect to hour of day, time of day, weekday, week, month and/or season. Determining the appliance-time associations, for an appliance, can be considered as grouping of sufficiently close appliance usage time-stamps, when that appliance has been recorded as active or operational, to form classes or clusters. The clusters or classes constructed will describe appliance-time associations while respective size of clusters, defined as the count of members in the cluster, will establish the relative strength for the clusters. The strength or size of the cluster will indicate how frequently and when a given appliance has been used by consumer, which indicate personal preferences. Therefore, discovery of appliance-time associations can be translated into clustering of appliances' operating time-stamps into brackets of time-spans, where each cluster belongs to an appliance with respective times-stamps (data points) as members of the cluster.

Cluster analysis is the process of creating groups from data points according to information extracted from the data, but without external intervention; i.e., unsupervised classification. This extracted information outlines the relationship among the data points

and acts as the base for classification to ensure data points within a cluster are closer to one another but distant from members of other clusters[83][84]. "Closer" and "Distant" are measures of association, defining how closely members of a cluster are related to one another. Hence, clustering analysis conducted over the input data, presented in tables (3.2) and (3.3), can generate clusters or classes defining natural associations of appliances with time where strength of cluster will define how strongly an appliance is associated with specific given time unit; i.e., hour of day, weekday, week, month or season. These appliance-time associations can identify peak load/energy consumption hours through usage concentration of appliances and/or explain the behavioral characteristics of occupants or consumers.

We extend one of the most widely used prototype-based partitional clustering techniques that is k-means cluster analysis to extract appliance-time associations. Our choice of clustering method is based upon popularity, ease of implementation and expected quicker run-times. K-means clustering technique can perform better than other unsupervised methods such as hierarchical clustering when processing a large dataset with low data dimensions. Time complexity for k-means can be linear in the number of data objects O(ndk) compared to quadratic O(n2 log n) for hierarchical clustering algorithms; where, n is the number of data objects, d is number of data dimensions, and k is number of clusters.

In the k-means cluster analysis, the prototype of a cluster is defined by its centriod, which is mean of all the member data points. In this approach clusters are formed from non-overlapping distinct groups; i.e., each member data point belongs to one and only one group or cluster [83]. Moreover, we mine the data incrementally in a progressive manner, which we explain next.

5.1.1 k-means Cluster Analysis

We introduce preliminary background on k-means clustering based on [84] and [83]. For a dataset, DB, having n data points in Euclidean space. Partitional clustering distributes the data points from DB into k clusters, $C_1, C_2, ..., C_k$, having centroids $c_1, c_2, ..., c_k$ such that $C_i \subset DB$, $C_i \cap C_j = \emptyset$ and $c_i \neq c_j$ for$(1 \leq i, j \leq k)$. An objective function based on Euclidean distance, equation (5.2), is used to measure the cohesion among data points, which reflects the quality of the cluster. This objective function is the sum of the squared error (SSE), define in equation (5.1), and k-means algorithm seek to minimize the SSE.

$$SSE = \sum_{i=1}^{k} \sum_{d \in C_i} distance(d, c_i)^2 \tag{5.1}$$

$$distance(x, y) = \sqrt{\sum_{j} (x_j - y_j)^2} \tag{5.2}$$

The k-means starts by selecting k data points from DB, where $k \leq n$, and form k clusters having centroids or cluster centers as the selected data points. Next, for each of the remaining data points d in DB, data point is assigned to a cluster having least Euclidean distance from its centroid (c_i); i.e., $distance(d, c_i)$. After a new data point is assigned revised, the cluster centroids are determined by computing the cluster centers for the clusters, and k-means algorithm repetitively refines the cluster composition to reduce intra-cluster dissimilarities by reassigning the data points until clusters are balanced; i.e., no reassignment is possible, while evaluating the cluster quality by computing the sum of the squared error (SSE) covering all the data points in the cluster from its centroid.

5.1.2 Optimal k - determine k using $Silhouette\ coefficient$

We make use of $silhouette\ coefficient$ calculated based on the $Eluclidean$ distance to determine the optimal number clusters k, while assessing the quality of clustering by analyzing intra-cluster cohesion and inter-cluster separation of data points among clusters. $Silhouette\ coefficient$ measures the degree of similarities and dissimilarities to indicates "How well clusters are formed?" $Silhouette\ coefficient$ can be computed as defined in equations (5.3, 5.4, 5.5, 5.6, and 5.7)[85].

- Compute a_j as average distance of d_j to all other data points in cluster C_i

$$a_j = \text{average}\{distance(d_j, d_i)\} \tag{5.3}$$
$$where, d_i = (d_1, d_2, ..., d_n); d_i \neq d_j$$

- Compute Average distance of d_j to all other data points in clusters C_i, having $i \neq j$; Determine b_j = $minimum(b_j)$ across all the clusters except C_i.

$$b_j = \text{average}\{distance(d_j, d_{i/C_x})\} \tag{5.4}$$
$$where, d_i = (d_1, d_2, ..., d_n);$$
$$and, C_x = (C_1, C_2, ..., C_n); C_x \neq C_i$$

- Compute $Silhouette\ coefficient$ for d_j

$$s_{d_j} = \frac{(b_j - a_j)}{maximum(a_j, b_j)} \tag{5.5}$$

- Compute $Silhouette\ coefficient$ for cluster C_i

$$s_{C_i} = average(s_{d_j})\ for\ j = d_1..d_n \tag{5.6}$$

- Compute $Silhouette\ coefficient$ for clustering, having k clusters

$$s_k = average(s_{C_i})\ for\ i = 1..k \tag{5.7}$$

$Silhouette\ coefficient$ can range from -1 to 1, where a negative value indicates misfit as the average distance of data point d_i to data points in the cluster C_i (a_i) is greater than the average distance d_i to data points in a cluster other than C_i (b_i) , and a positive number indicates better-fit clusters. Overall quality of a cluster can be assessed by computing average $Silhouette\ coefficient$ by computing the average of $Silhouette\ coefficient$ ($Silhouette\ width$) for all the member data points for the cluster, as shown in equation (5.6). Similarly, an average $Silhouette\ coefficient$ ($Silhouette\ width$) can be calculated for complete clustering by obtaining the average of $Silhouette\ coefficient$ for all the member data points across all the clusters, which is identified as $Global\ Silhouette\ coefficient$, as shown in equation (5.7) [85].

Finally, to determine optimal $k \in 1, 2, 3...n$, where n is the unique set of data points in database/dataset, the process of analyzing the quality of clusters formed is repeated while computing $Silhouette\ coefficient$ ($Silhouette\ width$) and k is chosen having maximum $Silhouette\ width$. Additionally, we use the approach proposed by [86] to efficiently compute $Silhouette\ coefficient$ ($Silhouette\ width$).

5.1.3 Optimal One-Dimensional k-means Cluster Analysis via Dynamic Programming

With reference tothe clustering input database, presented in tables (3.2) and (3.3), we can deduce that we have a cluster analysis requirement for a single dimension data. We make use of dynamic programming algorithm for optimal one-dimensional k-means clustering proposed by [87], which ensures optimality and efficient runtime. We further extend the algorithm to achieve incremental data mining to discover Appliance-time associations.

Here, we provide the relevant background based on [87]. A one-dimension k-means cluster analysis can be viewed as grouping n data points $d_1, d_2, ...d_n$ into k clusters, while minimizing sum of the squared error (SSE), shown in equation (5.1). A sub-problem for the original dynamic programming problem can be defined as finding the minimum SSE of clustering $d_1, d_2, ...d_i$ data points into m clusters. The respective minimum SSE are stored in a Distance Matrix (DM) of size $[n+1, k+1]$, where $DM[i, m]$ records the minimum SSE for the stated sub-problem and $DM[n, k]$ provides the minimum SSE for the original problem. For a data point d_j in the m clusters, where d_j is the first data element of cluster m, the optimal solution (SSE) to the sub-problem is $DM[i, m]$; therefore, $DM[j-1, m-1]$ must be the optimal SSE for the first $j-1$ data points in $m-1$ clusters. This establishes following optimal substructure defined by equation (5.8).

$$\begin{aligned} DM[i,m] = \min_{m \le j \le i} \{ & DM[j-1, m-1] \\ & + dist(d_j, ..., d_i)\} \\ & 1 \le i \le n,\ 1 \le m \le k \end{aligned} \tag{5.8}$$

Where $dist(d_j, ..., d_i)$ is computed SSE for $d_j, ..., d_i$ from their centroid/cluster center, and $DM[i, m] = 0$, for $m = 0$ or $i = 0$. $dist(d_j, ..., d_i)$ is computed iteratively from $dist(d_{j+1}, ..., d_i)$ as shown in equation(5.9) .

$$dist(d_j, ..., d_i) = dist(d_j, ..., d_{i-1}) + \frac{i-1}{i}(x_i - \mu_{i-1})^2 \tag{5.9}$$

$$\mu_i = \frac{x_i + (i-1)\mu_{i-1}}{i} \tag{5.10}$$

where, μ_{i-1} is the mean of the first $(i-1)$ data elements.

A backtrack matrix BM is maintained, of size $[n, k]$, to record the starting index of the first data element of respective cluster. Backtrack matrix is used to extract cluster members by determining the starting and ending indices for the corresponding cluster and retrieving data members from the original dataset. Equation (5.11) captures this notion.

$$\begin{aligned} BM[i, m] = \underset{m \le j \le i}{\operatorname{argmin}} \{ & DM[j-1, m-1] \\ & + dist(d_j, ..., d_i)\} \\ & 1 \le i \le n,\ 1 \le m \le k \end{aligned} \tag{5.11}$$

5.1.4 Incremental Mining - Cluster Analysis

We achieve incremental progressive clustering by merging clusters and/or adding new clusters extracted during each successive mining operation into a persistent database in the Database Management System. Discovered clusters database records all the relevant cluster parameters and information including centroid, SSE, $Silhouette\ coefficient/width$,

and data points and their distance from centroid. This enables the easy addition of new data points or clusters, while computing cluster parameters with respect to the newly added data points and updating the information in the database accordingly. Considering, the conversion of the time series data into a 30 minutes time-resolution source data during the data preparation phase it was determined that this time-resolution unit is sufficient to capture vital information regarding consumer energy consumption decision patterns. The cluster analysis for hour of day done on this source data will result in clusters created with a separation between cluster' centroids as multiples of such time-resolution, which is 30 minutes in the current case. This time-resolution is identified as $permissible\ centroid\ distance$. Whereas, cluster analysis on the other bases such as time-of-day, weekday, week , month and seasons have natural segmentation. With this operation we achieve more exclusive homogeneity and separation among clusters.

Upon completion of mining on the incremental quantum of data or current step, newly discovered clusters are matched against the existing clusters in the database to determine the closest cluster(s) to merge with, having centroid within the distance of $permissible\ centroid\ distance$ from the new cluster. If there exists no cluster in the database, which is closer to the new cluster satisfying the permissible centroid distance constraint, the new cluster is added/stored into the discovered clusters database with all the accompanying parameters and information. But, in case of success, data points from the new cluster will be added to the search(ed) cluster(s) while evaluating the quality of the final cluster by computing the $Silhouette\ coefficient/width$. Data points from the new cluster are picked according to increasing order of distance from the centroid. The most stable cluster configuration, having maximum $Silhouette\ coefficient/width$, are saved to the database. Algorithm (7) outlines the incremental cluster analysis using one-dimensional

Algorithm 7 Incremental Clustering: k-means

Require: Transaction database DB, permissible centroid distance between clusters $permissible\ centroid\ distance$ = 30

Ensure: Incremental clustering, clusters stored in discovered clusters database CL_DB

1: **for all** 24 hour quantum transaction data db_{24} in DB **do**
2: Determine optimal k for data quantum db_{24} by computing the $silhouette\ coefficient(width)$ for clustering
3: Discover k clusters CL_{24} in db_{24} using one-dimension k-means clustering via dynamic programming, while capturing SSE, Silhouette coefficient (width), and data points with distance from centroid
4: **for all** Clusters in CL_{24} **do**
5: Search for the closest cluster CL_DB, having its centroid within a distance of $permissible\ centroid\ distance$
6: **if** Cluster(s) found **then**
7: Merge clusters while evaluating the quality of the cluster by computing the $silhouette\ coefficient(width)$
8: **else**
9: Add the new cluster with respective parameters and information
10: **end if**
11: **end for**
12: **end for**

k-means clustering through dynamic programming and results is presented in table (5.1).

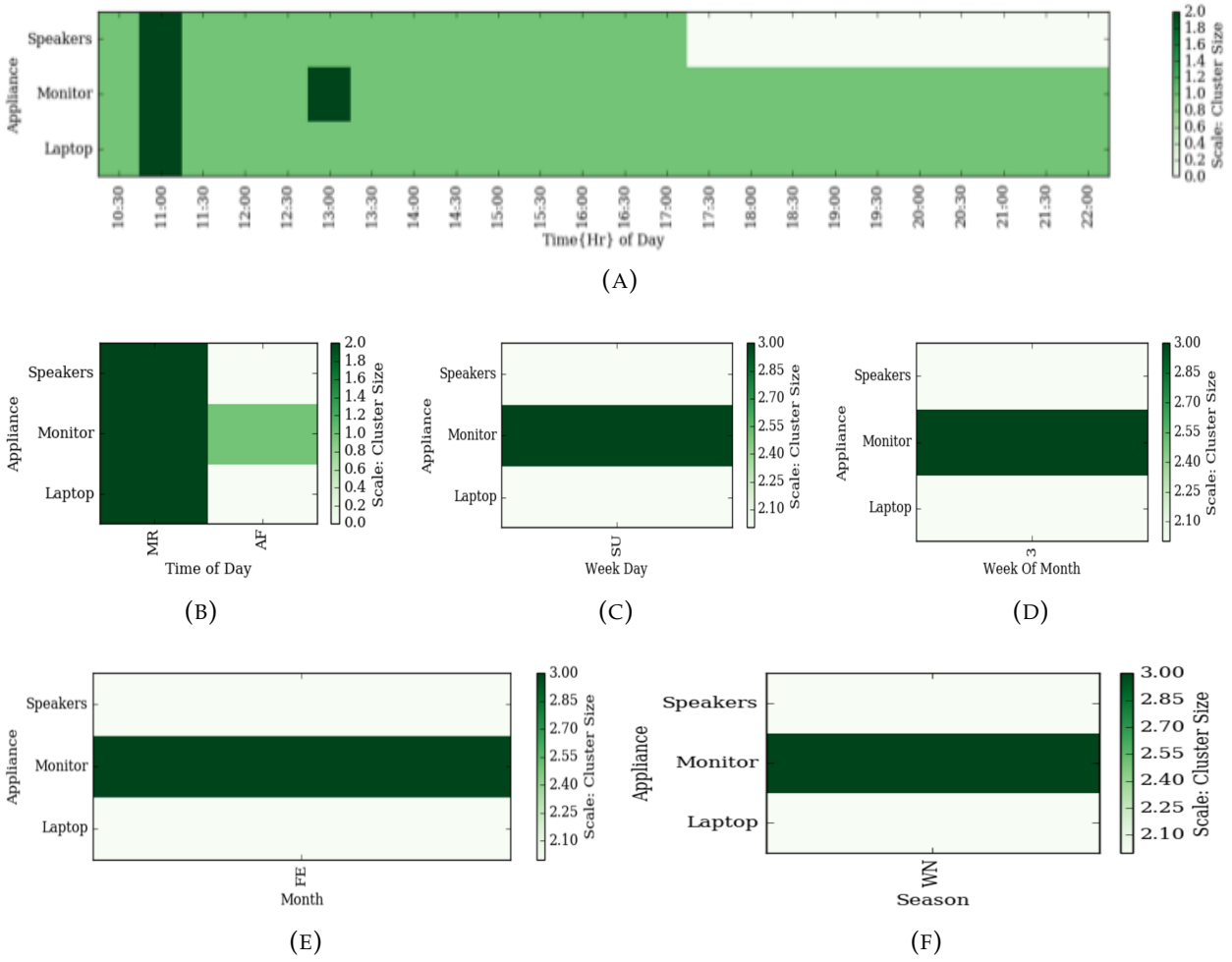
TABLE 5.1: Cluster Analysis: Clusters Discovered Database

Appliance	Cluster ID	Size	Centroid	SSE	Distance From Centroid
2	1	113	630	0	0
2	2	118	660	0	0
2	3	120	690	0	0
:	:	:	:	:	:
2	14	154	1020	0	0
2	15	151	1050	0	0
:	:	:	:	:	:
2	40	10	1380	0	0
2	41	8	1410	0	0
2	42	7	0	0	0
:	:	:	:	:	:
2	48	51	150	0	0
2 = Laptop					

5.2 Results

5.2.1 Results - Summary

In-depth incremental cluster analysis were conducted on the energy consumption data from five houses in the dataset UK-Dale [8] and one house from the dataset AMPds2[9], and synthetic dataset to inspect intermediate and final results. The cluster analysis results from three houses representative of our findings are presented. Figures (5.1-5.8), (5.9), (5.10), and (5.11) illustrates appliance-time associations for (A) the hour of the day, (B) the time of the day, (C) the weekday, (D) the week, (E) the month, and (F) the season. Figures (5.1-5.8) also show incremental appliance-time association construction for House 2. Figure (5.12) outlines the number of clusters formed during the incremental data mining process. A detailed discussion of the results is presented in next section (5.2.2).



(A)

(B)

(C)

(D)

(E)

(F)

FIGURE 5.1: House 2: Appliance-Time Associations [Upto 1 day Training Dataset]

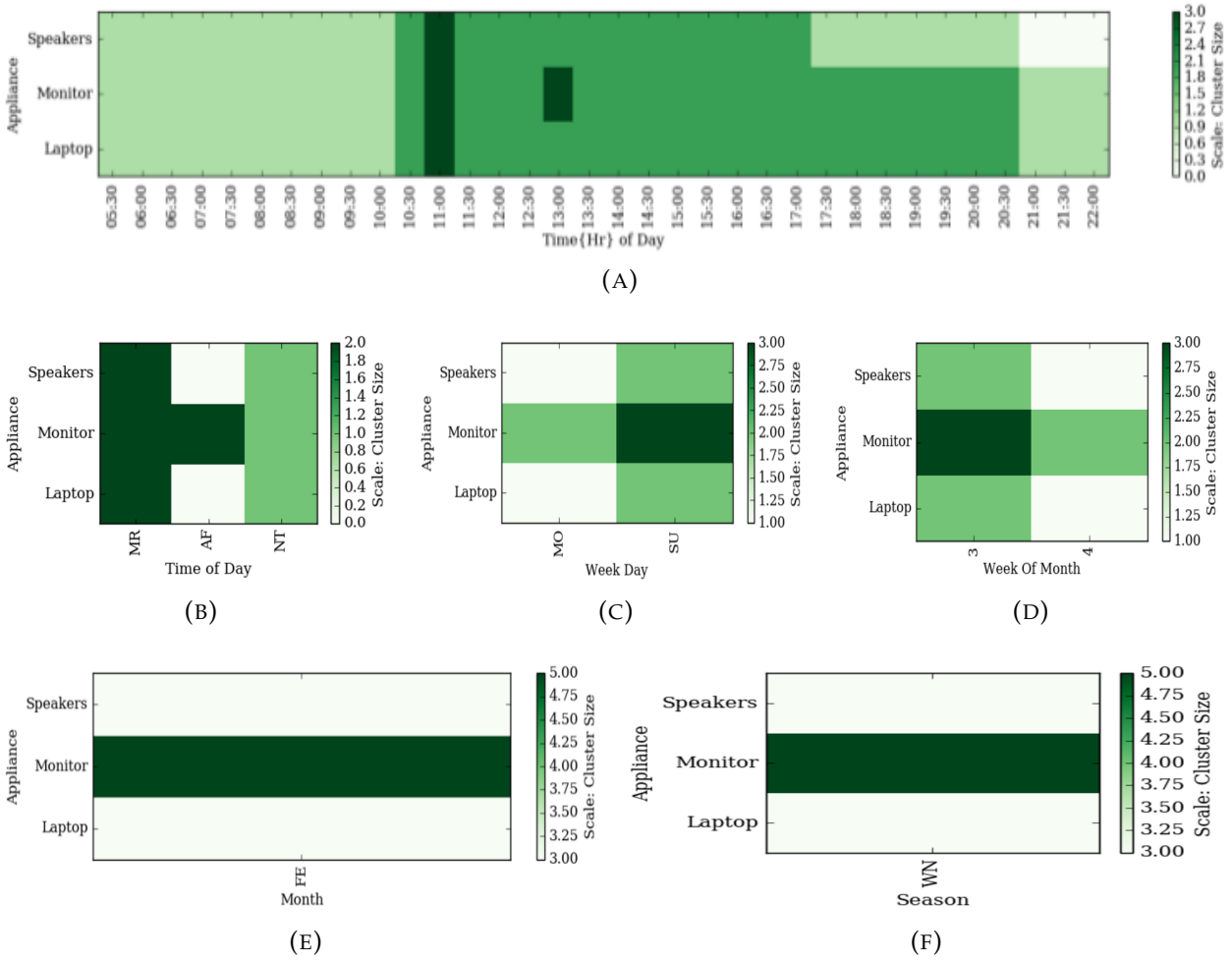


FIGURE 5.2: House 2: Appliance-Time Associations [Upto 2 day Training Dataset]

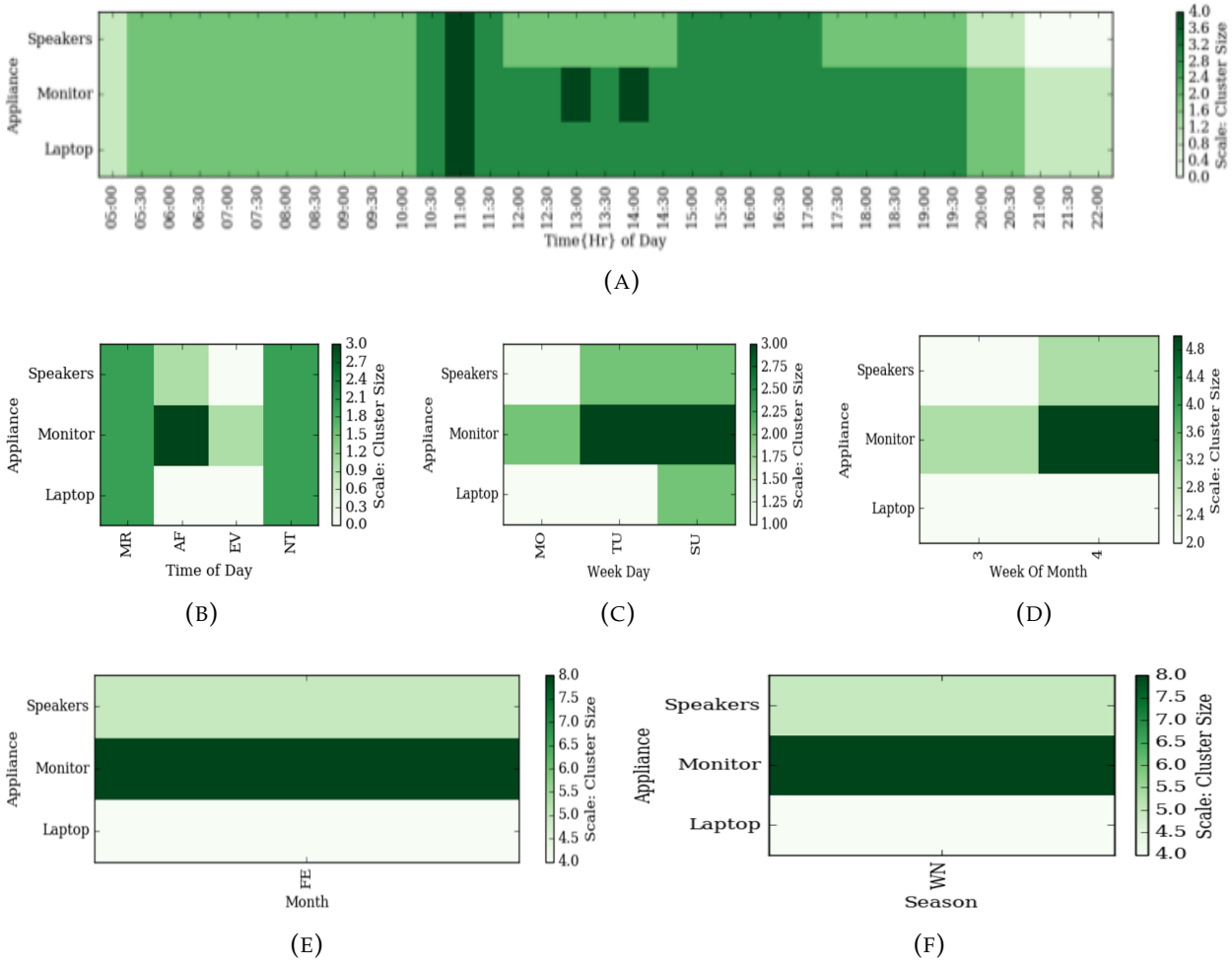


FIGURE 5.3: House 2: Appliance-Time Associations [Upto 3 day Training Dataset]

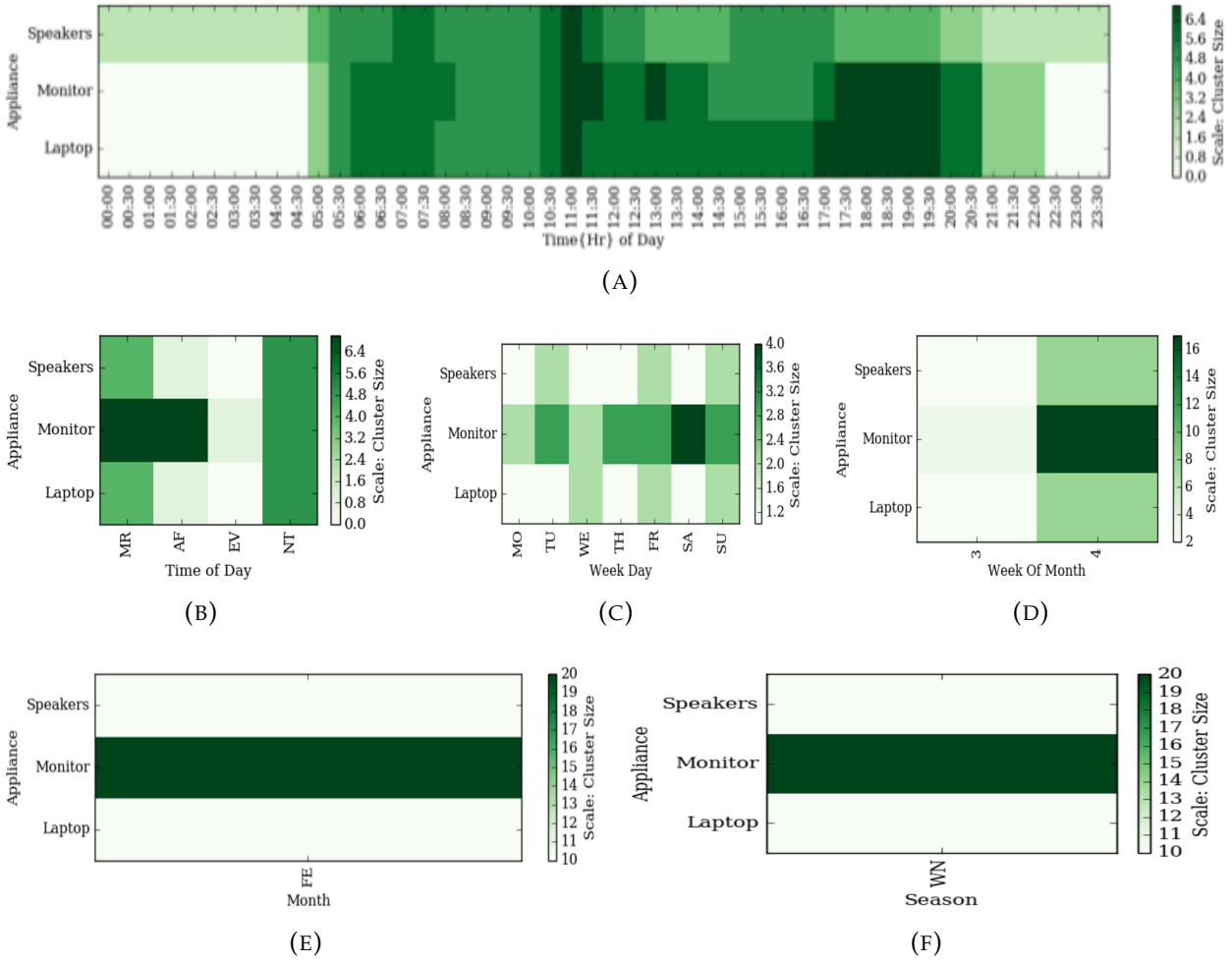


FIGURE 5.4: House 2: Appliance-Time Associations [Upto 7 day Training Dataset]

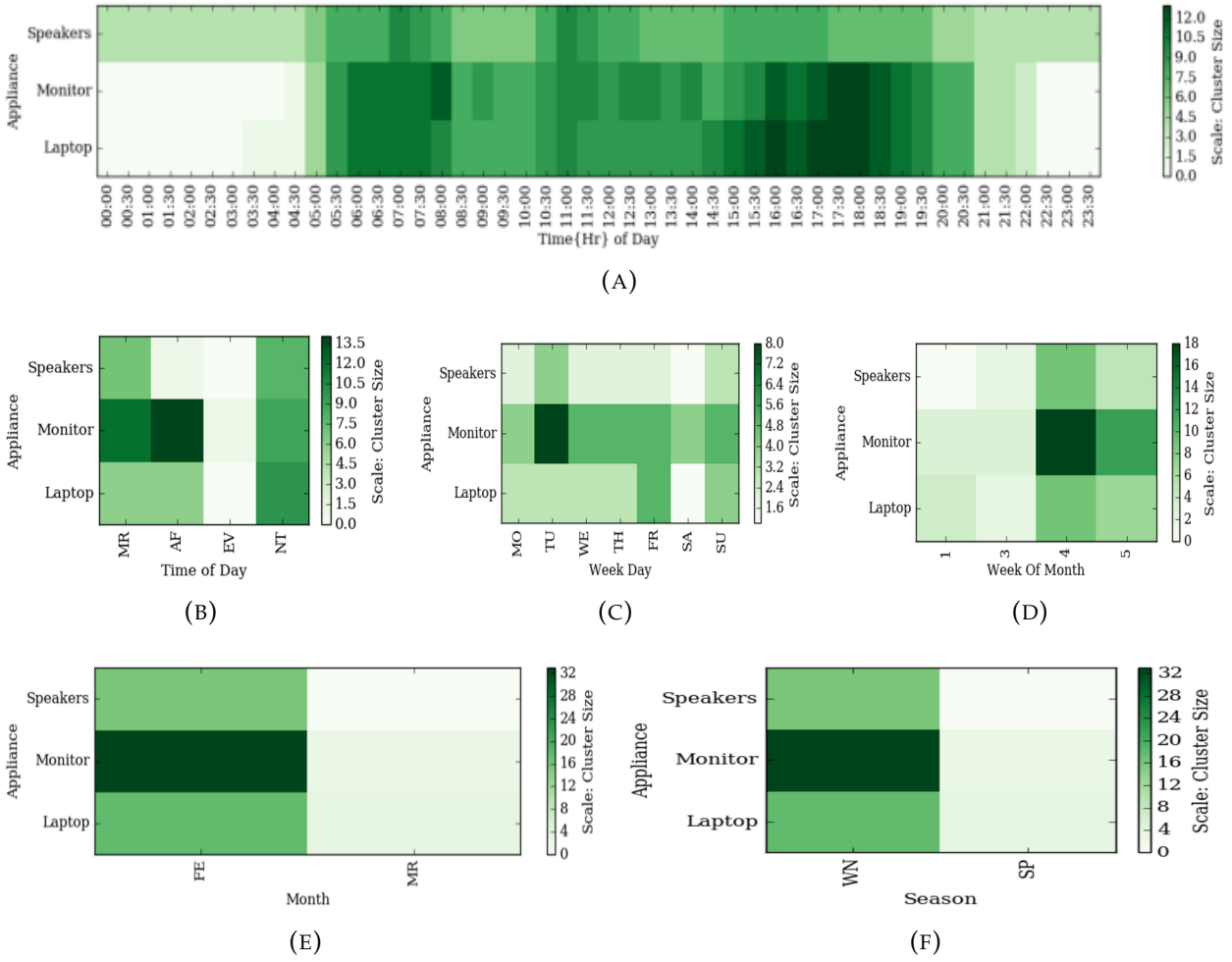


FIGURE 5.5: House 2: Appliance-Time Associations [Upto 15 day Training Dataset]

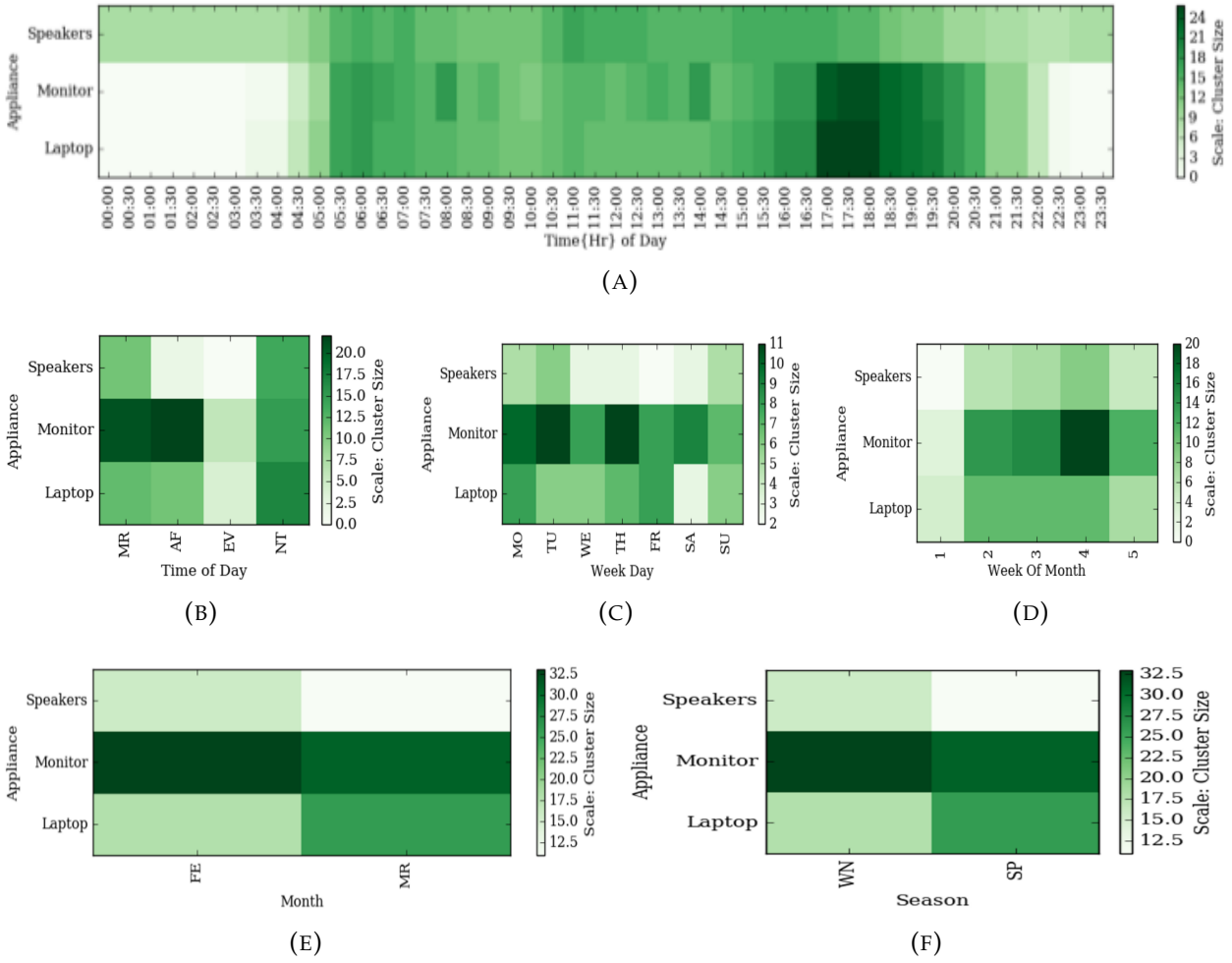


FIGURE 5.6: House 2: Appliance-Time Associations [Upto 30 day Training Dataset]

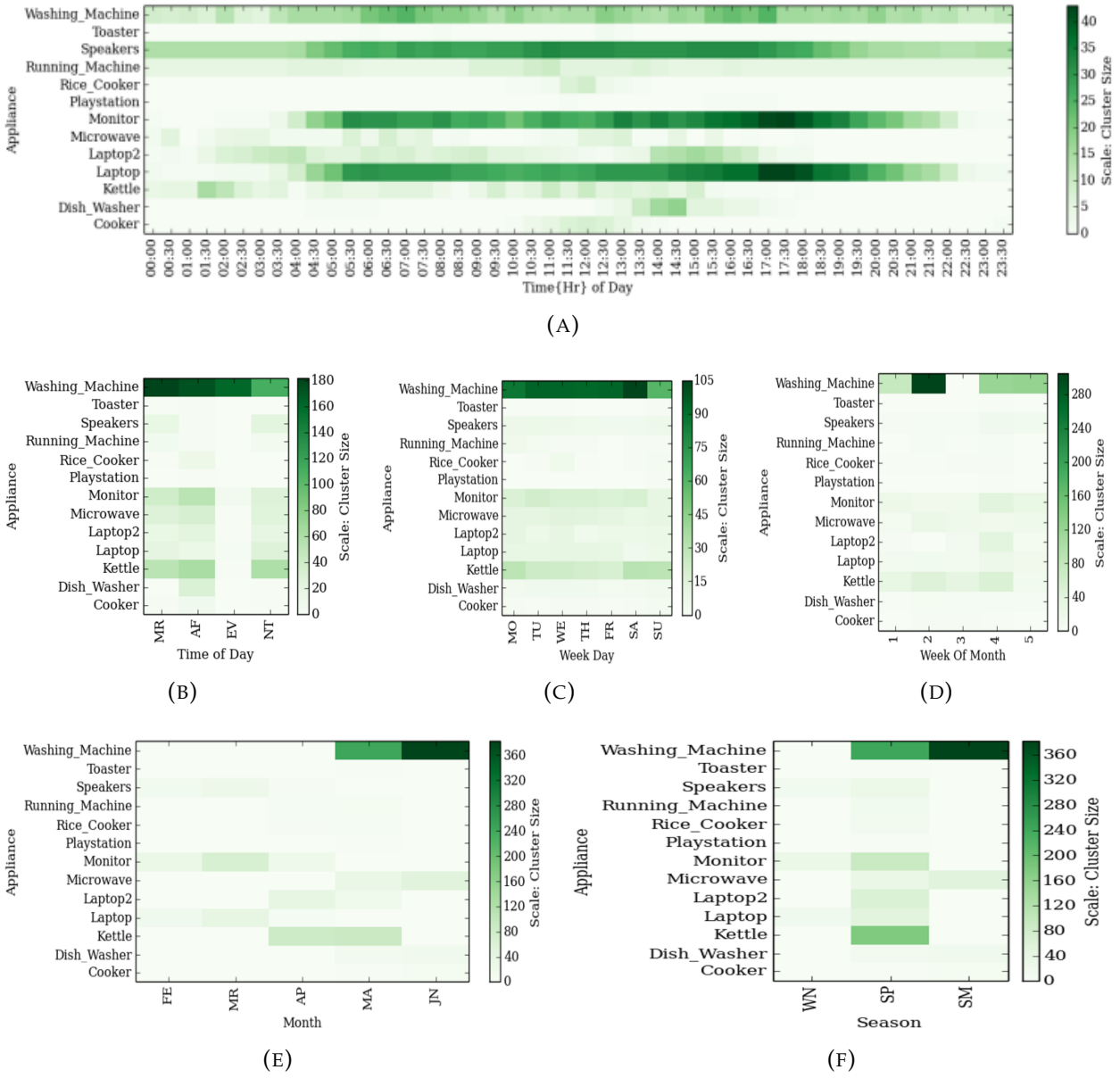


FIGURE 5.7: House 2: Appliance-Time Associations [25% Training Dataset]

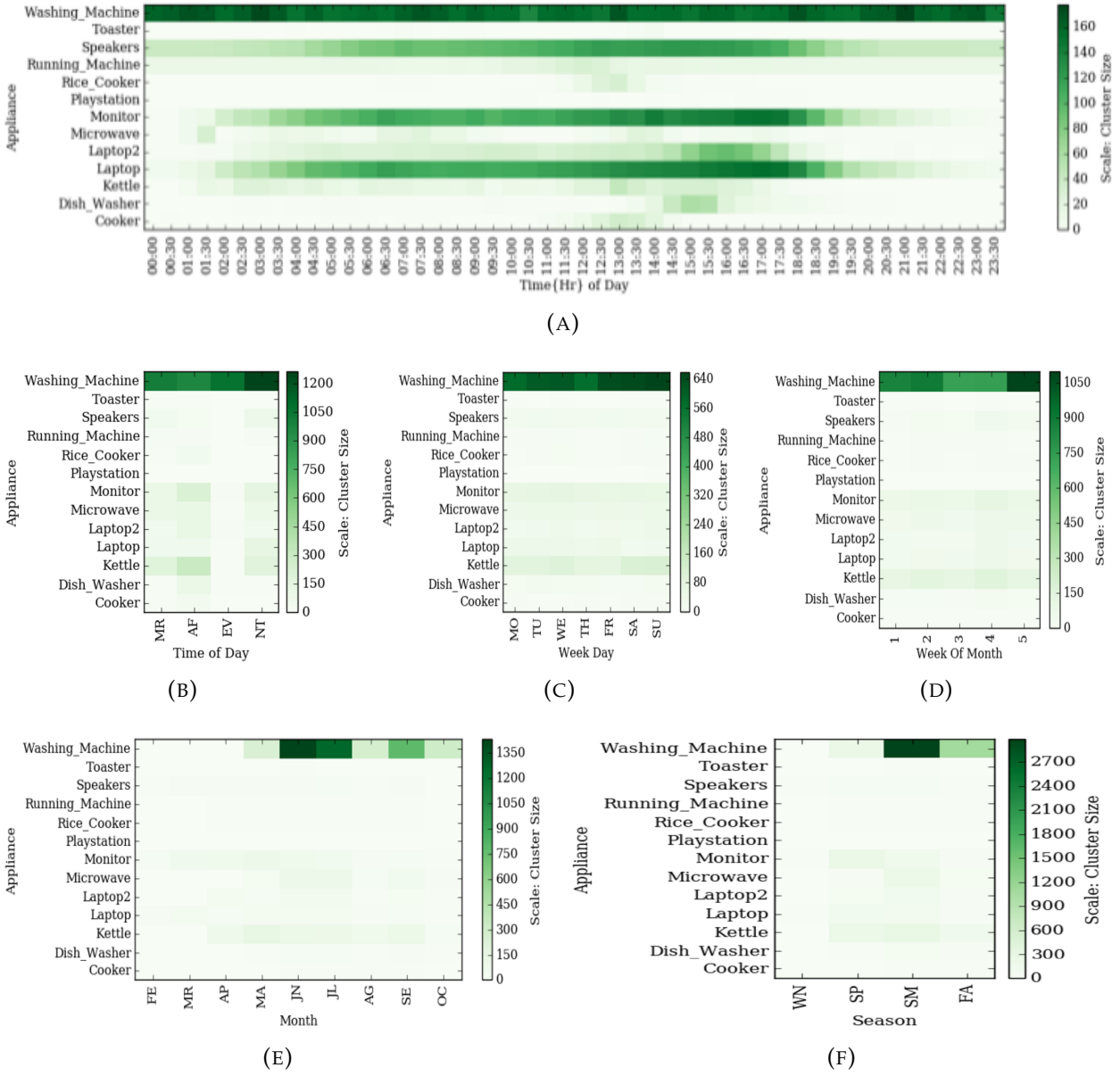


FIGURE 5.8: House 2: Appliance-Time Associations [Full Dataset]

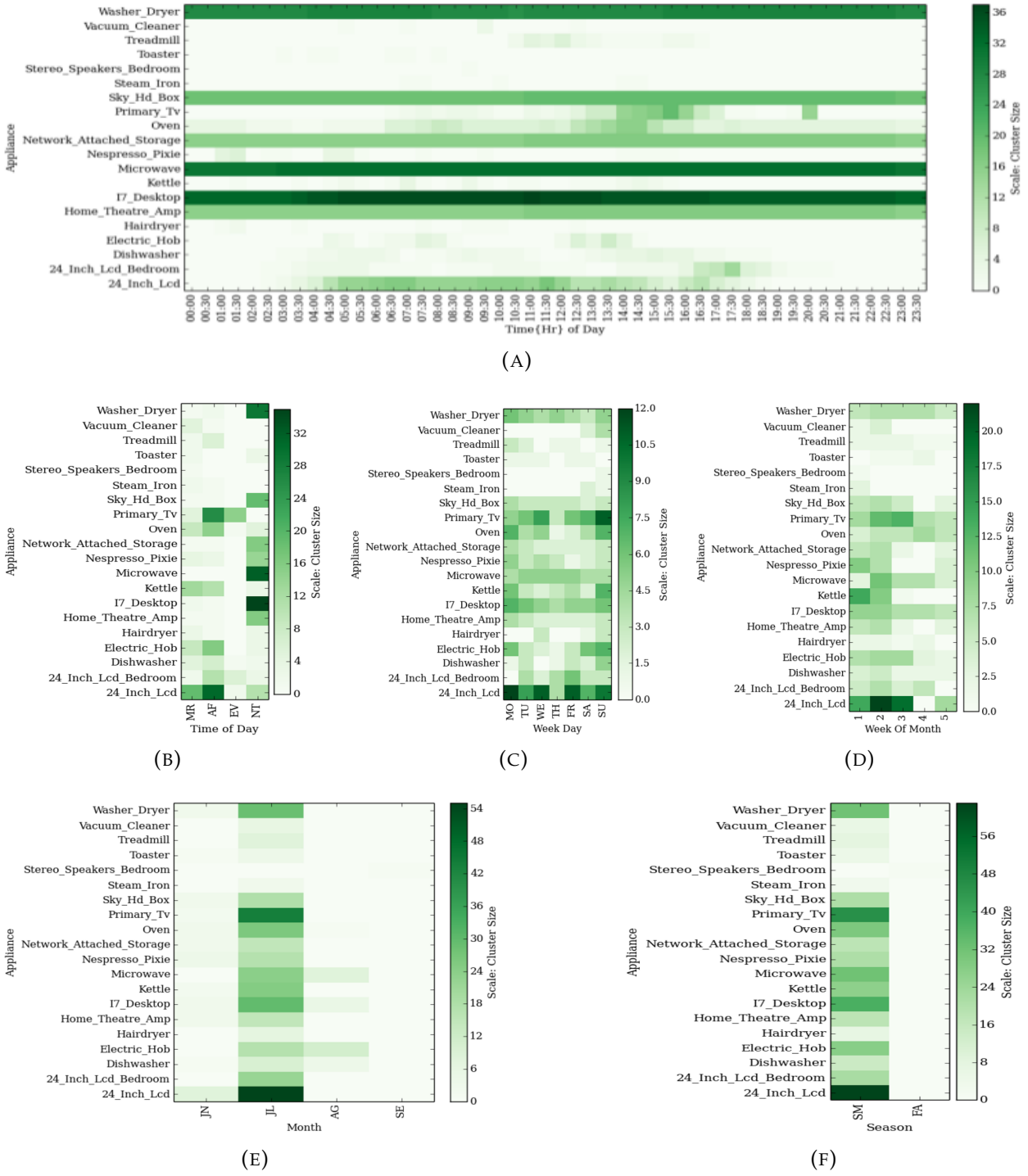


FIGURE 5.9: House 5: Appliance-Time Associations [25% Training Dataset]

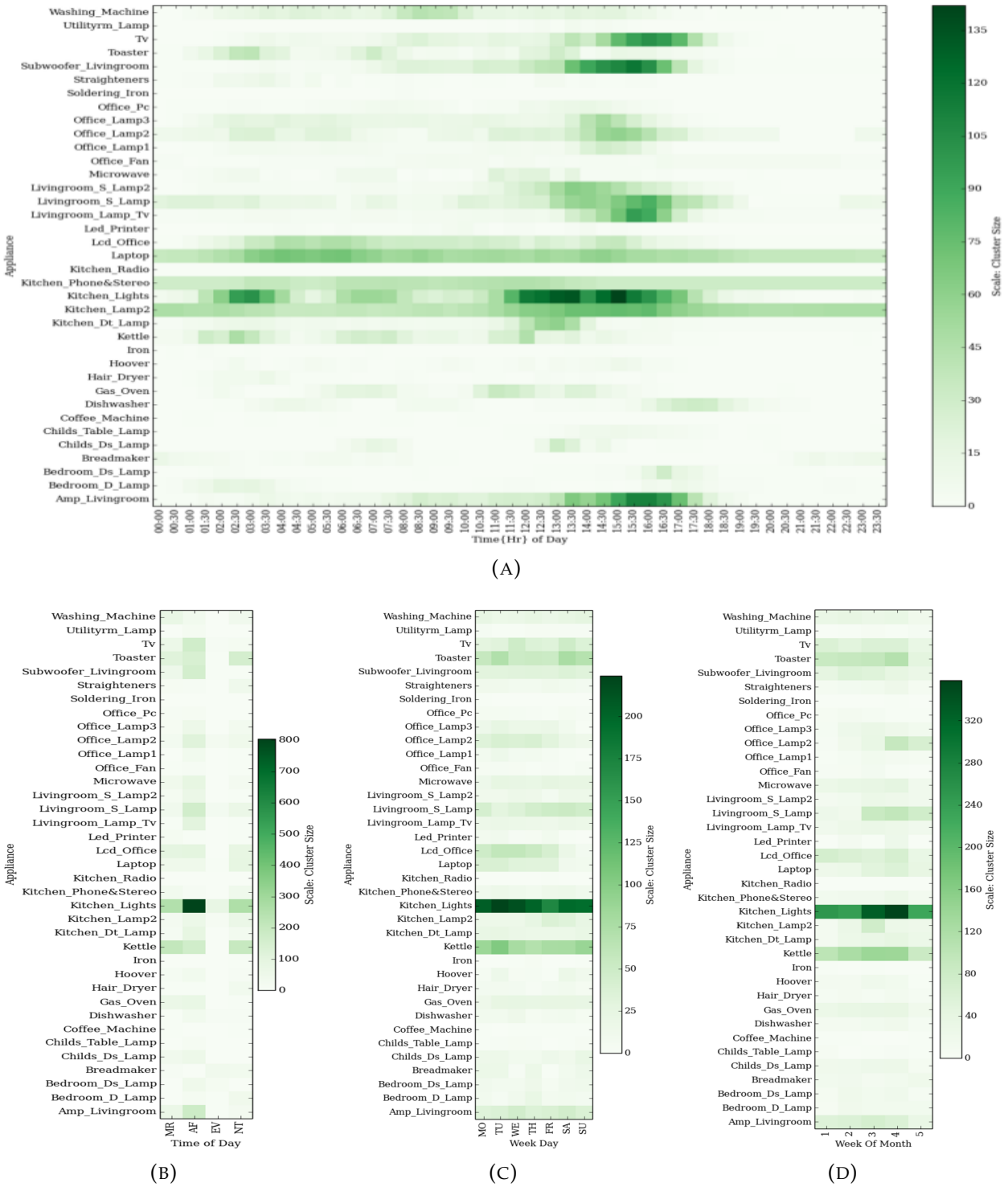
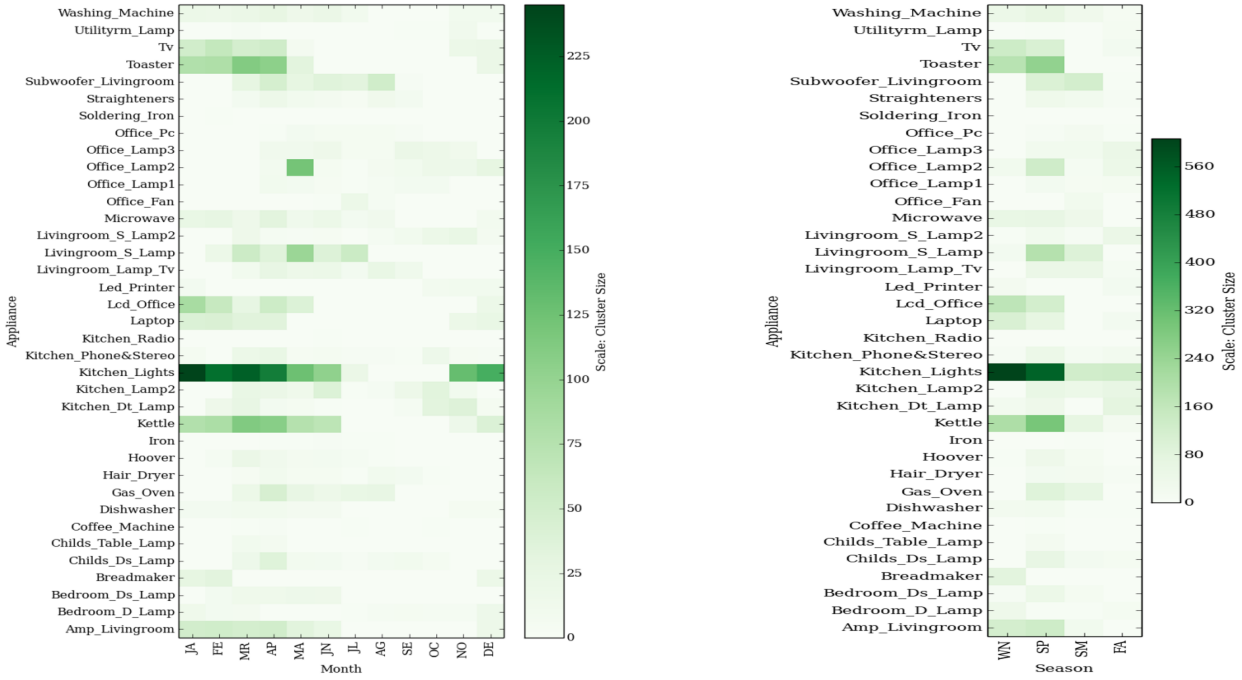
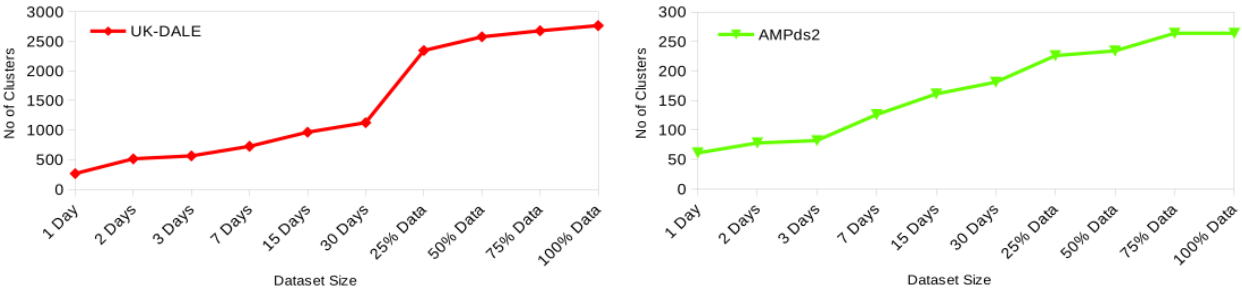


FIGURE 5.10: House 1: Appliance-Time Associations [25% Training Dataset] Part I



(A) (B)

FIGURE 5.11: House 1: Appliance-Time Associations [25% Training Dataset] Part II



(A) Dataset UK-Dale[8] (B) Dataset AMPds2[9]

FIGURE 5.12: Number of Clusters Discovered Vs Dataset Size Mined/Processed

5.2.2 Results - Analysis and Discussion

For House 2 (5.7), it was noticed that Laptop and Monitor are two appliances which were used simultaneous with the highest concentration of use during 16:45 - 18:15 and increased usage during the fourth week of the month and the spring with less frequent use on the weekends. The Washing Machine was used all day, with major usage concentration noted from 05:45 to 07:15 and 15:15 to 17:15 and increased usage observed on Saturdays, the second week of the month, and during the summer.

For House 1 ((5.10) and (5.11)), it was observed that the Kitchen Lights had a higher usage frequency between 02:15 to 03:45, and 11:15 to 17:15 with the highest usage frequency on Tuesdays and increased usage during weeks three and four of the month, and during Winter and Spring, although lower usage was observed on Fridays. The Laptop was under use throughout the day with increased usage frequency occurring from 02:45 to 06:45, during the fourth week of the month, and during Winter and Spring, but usage was reduced on the weekends.

House 5 (5.9), it was detected that personal computer was used all the day, with increased usage from 04:45 to 16:45, on Mondays, Tuesdays and Sundays, during the first and second weeks of month, and during Summer. The Washer and Dryer were used at a constant frequency, with reduced usage on Saturdays, and during the first and the last weeks of month. The Microwave was used with increased frequency on Tuesdays, Wednesdays, Thursdays and Fridays, but with reduced frequency during the first and the last weeks of the month. Also, Audio equipment was used regularly throughout the day with increased usage on Mondays and during the first two weeks of the month.

Figure (5.12) shows the incremental discovery of appliance-time associations in

form of clusters during the incremental data mining process, which further confirms the discovery of new appliance-time associations that are representative of energy consumption behavioral changes taking place over a period of time.

Appliance-time associations revealed by the cluster analysis strongly support the energy consumption analysis outcomes discussed in Chapter (4): low power rating appliances such as Laptop, Audio devices, Lights superseded high power appliances in terms of contributing towards household energy bills owing to their extended usage duration.

Based on these results, the varying effects of time, days, and seasons on appliance usage presents strong proof for the impact of consumer behavior on household energy consumption patterns translating into household energy consumption, which strengthen support to the presented argument in this research. The outcome of the evaluation thoroughly reinforced the conjecture of unwavering impact of human behavior on energy consumption patterns at a household learned from appliance-time associations and appliance-appliance associations. Therefore, appliance association with time and other appliances can reveal consumer energy consumption decision patterns which are direct reflections of personal energy consumption lifestyle preferences and anticipated levels comfort.

Chapter 6: Multiple Appliance Usage Prediction and Energy Consumption Forecast

6.1 Bayesian Network for Multiple Appliance Usage Prediction and Household Energy Forecast

we utilize a Bayesian Network to predict the use of multiple appliances at some point in the future. Bayesian network is a directed acyclic graph, where nodes represent random variables and edges indicate their probabilistic dependencies. Each node or variable is independent of its non-descendants and accompanied by its local conditional probability distributions in the form of a node probability table, which facilitates the computation of joint conditional probability distribution for the model [88][89][86]. Therefore, the local probability distributions furnish quantitative probabilities that can be multiplied according to qualitative independencies described by the structure of Bayesian Network to obtain the joint probability distributions for the model. Bayesian network has advantages such as - ability to effectively mitigate missing data, learn relationships, and make use of historical facts and observations while avoiding over-fitting of data [90][91][92]. A Bayesian network is defined by the probabilistic distribution presented in equation (6.1) [93] [94] [95].

$$p(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} p(x_i|parents(x_i)) \tag{6.1}$$

6.1.1 Probabilistic Prediction Model

We construct our model based on Bayesian network, having seven nodes representing probabilities for appliances-to-time associations in terms of hour of day (00:00 - 23:59), time of day (Morning, Afternoon, Evening, Night), weekday, week, month, season (Winter, Summer/Spring, Fall), and appliance-appliance associations. The resulting Bayesian network has a very simple topology, with known structure and full observability, comprising only one level of input evidence nodes, accompanied by respective unconditional probabilities, converging to one output node. The equation (6.2) and figure (6.1) presents the posterior probability or marginal distribution and network structure for the proposed prediction model respectively.

$$\begin{aligned} p(.) = p(Hour) \times p(Time\ of\ day) \\ \times p(Weekday) \times p(Week) \\ \times p(Month) \times p(Season) \end{aligned} \tag{6.2}$$

We use the output of the earlier two phase's; i.e., frequent pattern mining and cluster analysis to train the model, which effectively incrementally learns the prior information through progressive mining. First set of inputs comes from frequent pattern mining outcomes; where marginal distribution for the appliance-appliance associations are computed at a global level by using $support$ for an itemset, which is probability of itemset in the transaction database, as shown in table (6.1). The calculated marginal distributions determines the probabilities of appliances being concurrently active.

The next set of inputs comes from cluster analysis outcomes; where marginal

TABLE 6.1: Frequent Patterns: Frequent Patterns Discovered Database

Frequent Pattern	Absolute Support (S)	Database Size (D)	Support or Probability $P = S/D$
'2 3'	3939	7899	0.4986
'2 4'	2840	7899	0.3595
'2 3 4'	2649	7899	0.3353
2 = Laptop, 3 = Monitor, 4 = Speakers			

distribution for appliance-time associations are computed at a global level using the cluster configurations for a given unit of time, which is cluster center or centroid; i.e., hour, time of day, weekday, week, month, and season, as shown in table (6.2). The calculated marginal distribution determines the probabilities of appliances being active during the period identified by the centroid.

TABLE 6.2: Cluster Analysis: Cluster Marginal Distribution

Appliance	Cluster ID	Size (S)	Probability (P) $P_i = S_i / \sum S_i$
2	40	10	0.0414
3	38	5	0.0207
4	37	41	0.1701
10	37	15	0.0622
12	6	168	0.6970
15	40	2	0.0082
2 = Laptop, 3 = Monitor, 4 = Speakers			
10 = Running Machine, 12= Washing Machine, 15 = Microwave			

Further, table (6.3) represents a sample of the training data with marginal distribution for the various appliances and the probabilities for appliances to be active during the period, for the node variable/parameter vector. The probabilities are computed from the clusters, formed and continuously updated during the mining operation.The respective cluster strength or size determines the relative probability for the individual appliance. Additionally, appliance-appliance association, the outcome of the frequent pattern mining, computes the probabilities for the appliances to operate or be active concurrently. Therefore, the model uses top-down reasoning to determine and predict active appliances,

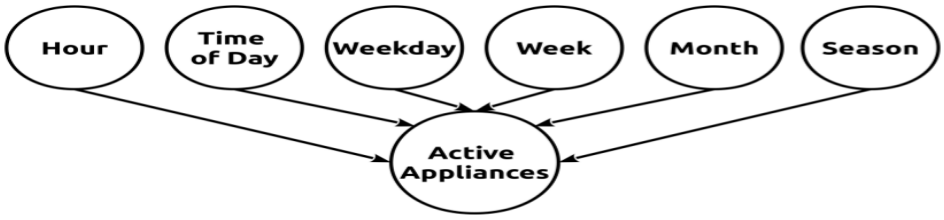


FIGURE 6.1: Bayesian prediction model: seven input evidence nodes

operating simultaneously, using the historical evidences from cluster analysis (Appliance-time association) and frequent pattern mining (appliance-appliance association).

Furthermore, appliance prediction results establish the foundation for household energy consumption forecast; where, average appliance load and average appliance usage duration are extracted from the historical raw time-series data at the respective time resolution that is hour, weekday, week, month, or season level. In this sense, model is capable of predicting energy consumption for a defined time in the future ranging from the next hour up-to 24 hours (short-term) and days, weeks, months, or seasons (long-term).

Moreover, for the purpose of evaluation and comparion of results and accuracy of the proposed model, we make use of a multi-label SVM classifier, based on Binary Relevance (an One-vs-All (OvA) approach) algorithm as problem transformation method [96][97], to predict concurrent operation of multiple appliances to compare results and accuracy of proposed model. The choice of SVM is influenced by its wide acceptance as preferred mechanism for prediction in the area smart meters data analytics. We provide a brief explanation on SVM multi-label classifier in next section; i.e., section (6.2), and then present evaluation results and analysis of results in section (6.3).

TABLE 6.3: Node Probability Table - Marginal Distribution for Appliances

	Cluster Center/Centroid				
Appliance	00:00	00:30	01:00	01:30	02:00
2	0.0608	0.0602	0.1257	0.1414	0.1404
3	0.0338	0.0361	0.1078	0.1212	0.1149
4	0.1351	0.1205	0.1198	0.1010	0.0936
8	0.0270	0.0422	0.0719	0.1010	0.0979
10	0.0878	0.0783	0.0778	0.0657	0.0553
11	0.0068	0.0181	0.0180	0.0404	0.0511
12	0.6014	0.5060	0.4731	0.3939	0.4043
13	0.0000	0.0000	0.0000	0.0051	0.0043
15	0.0473	0.1386	0.0060	0.0303	0.0383

2 = Laptop, 3 = Monitor, 4 = Speakers
8 = Kettle, 10 = Running Machine, 11= Laptop2
12= Washing Machine, 13 = Dishwasher, 15 = Microwave

6.2 SVM - Multi-label Classifier

Objective of the classification exercise is to uncover information which can be utilized to predict or assign class to an unknown input pattern, by examining the status of attributes for this input pattern. This supervised learning approach has two flavors that is single-label and multi-label classification, where classification task is to assign one label or two or more unique labels to the input instance respectively [98] [97]. Further, Binary Relevance (one-vs-all) is most commonly used problem transformation method; in which one classifier is trained for each class label over original dataset that is original problem is transformed into L problems, where L represents the number of labels, and one classifier is trained for each problem. Resultant labelset for each transformed problem contains same number labels with positive label for original class but negative otherwise. For the classification of an unknown input pattern, Binary Relevance produces unified labelset of the labels that were positively predicted by the L classifiers[97][99].

According to [98], [97], [99] and [100], multi-label classification can be defined as following:

For a $M - dimensional$ feature attributes input space $\mathcal{X}$, and applicable $L - dimensional$

$labelset\ \mathcal{L} = \{\lambda_i\ :\ i\ =\ 1\ \cdots\ L\}$ with label associations $Y \subseteq \mathcal{L}$ and having N example training set:

$$T = \{x^{(i)}, y^{(i)}\}\ for\ \ i = 1 \ldots N$$

$$= \begin{pmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_M^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots & x_M^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{(N)} & x_2^{(N)} & \cdots & x_M^{(N)} \end{pmatrix} \begin{pmatrix} y_1^{(1)} & y_2^{(1)} & \cdots & y_M^{(1)} \\ y_1^{(2)} & y_2^{(2)} & \cdots & y_M^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ y_1^{(N)} & y_2^{(N)} & \cdots & y_M^{(N)} \end{pmatrix}$$

$where,$

$x^{(1)} = [x_1, \cdots, x_M] \in \mathcal{X}$

$y^{(1)} = [y_1, \cdots, y_N] \in \mathcal{L}$

where, $y_i = \begin{cases} 1 & \text{if } i^{th} label \text{ is relevant} \\ 0 & \text{otherwise} \end{cases}$

Hence, an i^{th} lableset is determined as a binary vector,

$Y^{(i)} \subset \mathcal{L} \Longleftrightarrow y^{(i)} = [y_1, \cdots, y_n]$

Therefore, a Multi-label classifier can be define as,

$h : \mathcal{X} \to \mathcal{Y}$

Finally, prediction for an input instance $\tilde{x}$ is determined as,

$\tilde{y} = h(\tilde{x}) = [h^{(1)}(\tilde{x}), h^{(2)}(\tilde{x}), \cdots, h^{(L)}(\tilde{x})]$

6.3 Results

6.3.1 Results - Summary

Extensive prediction experiments were conducted to analyze the accuracy of the proposed model using two datasets UK-Dale [8] and AMPds2 [9] along with a synthetic dataset to

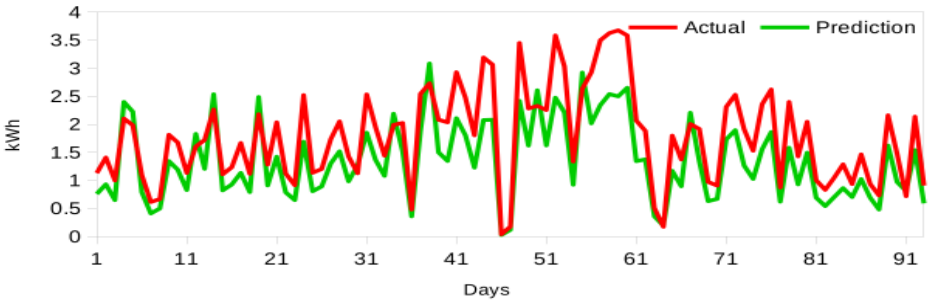
FIGURE 6.2: Prediction Accuracy: Proposed Model Vs SVM

analyze prediction accuracy of our proposed model. Results from the appliance-appliance and the appliance-time associations were used in the probabilistic prediction model to predict multiple concurrent operating appliances with great success. The prediction results for short term, long term predictions, and overall prediction, are provided for the three stages of the incremental data mining process (i.e. 25%, 50% and 75%, respectively). Figure (6.2) shows a comparison of the accuracy of the proposed model with the SVM at different stages of training. Figures (6.4) and (6.5) show the accuracy of the proposed model at the overall and premises level, respectively compared to the SVM. Additionally, figure (6.3) shows the accuracy of household energy consumption forecast at short and long term time frames. A detailed discussion of the results is presented in section (6.3.2).

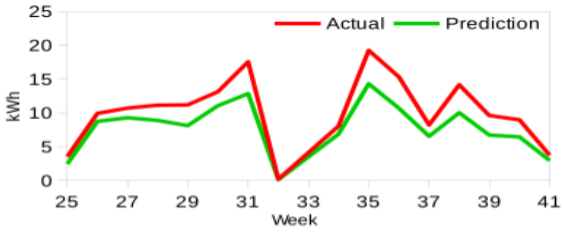
6.3.2 Results - Analysis and Discussion

The proposed model outperformed the SVM and attained a combined accuracy of 81.65 (25%), 85.90 (50%), 89.58 (75%) (figure (6.2), and table (6.4)) at each stage of training respectively, which support the hypothesis that incremental mining can identify variations induced by dissimilarities in household occupants' behavioral traits and can facilitates well informed energy consumption decision making at various levels. Subsequently, the results of multiple appliances predictions were applied to forecast expected household

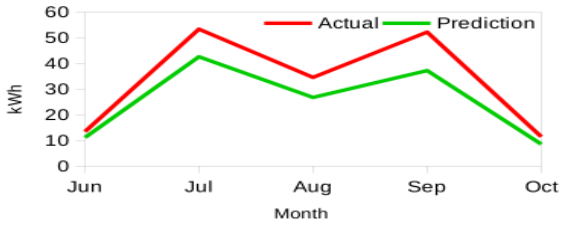
energy consumption. The accuracy achieved was 81.89%, 75.88%, 79.23%, 74.74%, and 72.81% (figure (6.3)) for short-term(hourly), long-term(daily, weekly, monthly, seasonal) energy consumption predictions, respectively.



(A) $Long - Term$:Day

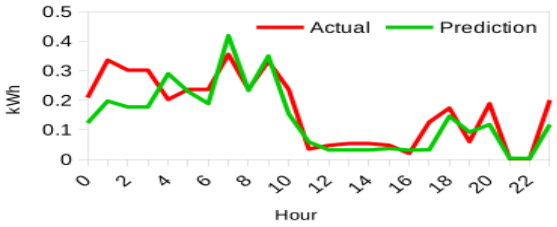


(B) $Long - Term$:Week



(C) $Long - Term$:Month

(D) $Long - Term$:Season



(E) $Short - Term$

FIGURE 6.3: House 2: Energy Consumption Prediction Vs Actual Energy Consumption

TABLE 6.4: Prediction : Model Accuracy, Precision, Recall

	Short Term			Long Term			Overall		
Model	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
25% Data as Training Data									
Proposed Model	85.19%	89.13%	75.93%	78.18%	76.09%	64.64%	81.65%	82.61%	69.72%
SVM	58.57%	75.93%	58.57%	58.06%	65.45%	58.06%	58.33%	70.64%	58.33%
50% Data as Training Data									
Proposed Model	89.41%	89.29%	88.24%	81.69%	81.43%	80.28%	85.90%	85.71%	84.62%
SVM	77.08%	77.08%	77.08%	63.95%	63.95%	63.95%	70.88%	70.88%	70.88%
75% Data as Training Data									
Proposed Model	91.67%	95.52%	88.89%	87.50%	88.41%	84.72%	89.58%	91.91%	86.81%
SVM	80.00%	92.75%	80.00%	78.67%	84.29%	78.67%	79.35%	88.49%	79.35%

TABLE 6.5: Prediction : Premise Level Accuracy, Precision, Recall @ 75% Data as Training Data

			Short Term			Long Term		
Dataset	Premises	Model	Accuracy	Precision	Recall	Accuracy	Precision	Recall
0	House_1	Proposed Model	81.82%	87.50%	63.64%	90.00%	87.50%	70.00%
		SVM	41.18%	70.00%	41.18%	50.00%	60.00%	50.00%
1	House_1	Proposed Model	90.00%	90.00%	90.00%	81.82%	81.82%	81.82%
		SVM	69.23%	90.00%	69.23%	69.23%	81.82%	69.23%
	House_2	Proposed Model	84.62%	100.00%	84.62%	90.91%	100.00%	90.91%
		SVM	91.67%	100.00%	91.67%	90.91%	100.00%	90.91%
	House_3	Proposed Model	87.50%	87.50%	87.50%	70.00%	70.00%	70.00%
		SVM	87.50%	87.50%	87.50%	70.00%	70.00%	70.00%
	House_4	Proposed Model	100.00%	100.00%	100.00%	90.00%	90.00%	90.00%
		SVM	100.00%	100.00%	100.00%	90.00%	90.00%	90.00%
	House_5	Proposed Model	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
		SVM	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
3	House_1	Proposed Model	91.67%	95.52%	88.89%	90.00%	90.00%	90.00%
		SVM	80.00%	80.00%	80.00%	90.00%	90.00%	90.00%

0 - Synthetic Dataset , 1 - UK-DALE Dataset [8], 3 - AMPds2 Dataset[9]

Chapter 7: Conclusion and Future Work

7.1 Conclusion and Future Work

This research demonstrated that the appliance associations, with time and other appliances, are direct reflections of occupant energy usage behavior patterns and reveal personal preferences representative of expected levels comfort in relation to the use of electrical appliances for energy consumers. Incremental appliance association discovery disclosed that these associations change over specific time increments, such as seasons, months, weeks, weekdays, time of day, and the hour of the day, due to the strong impact of occupants' changing personal preferences on energy consumption decision patterns, which eventually gets translated into household energy consumption. Therefore, it is critical to learn these variations at regular intervals to identify and record varying impact of consumers' behavior on energy consumption. These appliance associations must be indispensable input parameters to the energy saving and efficiency programs and related decision making processes that seek to ensure consumer confidence and facilitate improved participation for successful and persistent program results. From the perspective of demand management scheduling, the appliances contributing to peak power load should be considered first in any energy savings plan, but consumer preferences should also be adequately accounted to minimize inconveniences to household occupants' day-to-day energy usage lifestyle in order to gain and retain consumer confidence. Additionally, manually operated appliances, which characterize occupants' energy consumption behavior, should also be considered when devising energy savings plans to maximize consumer agreement, increase participation, and ensure benefits to consumers. These programs

would require active engagement of occupants, either through automated or manual control mechanisms, in personal energy management at household level.

This research presented unsupervised incremental and progressive data mining, to identify the impact of consumers' behavioral personal preferences on energy consumption decision patterns based on the appliance-appliance and appliance-time associations, along with a prediction model to predict multiple appliances usage and forecast household energy consumption. It is important to note that the proposed mechanism can be applied to any quantum of time size for incremental mining, although current research considered 24 hours as the most optimal selection of time-span for retrieving the underlying essential information of concern. Extensive experiments were performed on the proposed model using real-world context-rich smart meter datasets, and the results showed that the proposed system outperformed the SVM predictor.

For future work, we plan to devise a methodology to consider the appliance usage duration along with actual power load while extracting and constructing appliance associations (with time and other appliances) from energy consumption data to attain improved accuracy in representing consumers' energy consumption behavioral decision patterns. We also plan to refine the model and introduce distributed learning through big data mining from multiple houses in a near real-time manner. This will allow utilities and consumers to react to momentarily energy consumption changes and aid in improving smart grid energy saving programs.

References

[1] T. Yu, N. Chawla, and S. Simoff, "Computational intelligent data analysis for sustainable development", in, ser. Chapman & Hall/CRC Data Mining and Knowledge Discovery SERIES. Chapman and Hall/CRC, Apr. 2013, ch. Chapter 7: Data Analysis Challenges in the Future Energy Domain, pp. 181–242, ISBN: 9781439895948. [Online]. Available: https://books.google.ca/books?id=iKx4qyZPvCcC.

[2] Y. C. Chen, H. C. Hung, B. Y. Chiang, S. Y. Peng, and P. J. Chen, "Incrementally mining usage correlations among appliances in smart homes", in *Network-Based Information Systems (NBiS), 2015 18th International Conference on*, Sep. 2015, pp. 273–279. DOI: 10.1109/NBiS.2015.43.

[3] D. Schweizer, M. Zehnder, H. Wache, H. F. Witschel, D. Zanatta, and M. Rodriguez, "Using consumer behavior data to reduce energy consumption in smart homes: Applying machine learning to save energy without lowering comfort of inhabitants", in *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, Dec. 2015, pp. 1123–1129. DOI: 10.1109/ICMLA.2015.62.

[4] C. O. Adika and L. Wang, "Autonomous appliance scheduling for household energy management", *IEEE Transactions on Smart Grid*, vol. 5, no. 2, pp. 673–682, Mar. 2014, ISSN: 1949-3053. DOI: 10.1109/TSG.2013.2271427.

[5] B. Anca-Diana, G. Nigel, and M. Gareth, "Achieving energy efficiency through behaviour change: What does it take?", Tech. Rep., 2013. DOI: 10.2800/49941. [Online]. Available: http://www.eea.europa.eu/publications/achieving-energy-efficiency-through-behaviour.

[6] A. Zipperer, P. A. Aloise-Young, S. Suryanarayanan, R. Roche, L. Earle, D. Christensen, P. Bauleo, and D. Zimmerle, "Electric energy management in the smart

home: Perspectives on enabling technologies and consumer behavior", *Proceedings of the IEEE*, vol. 101, no. 11, pp. 2397–2408, Nov. 2013, ISSN: 0018-9219. DOI: 10.1109/JPROC.2013.2270172.

[7] S. Singh, A. Yassine, and S. Shirmohammadi, "Incremental mining of frequent power consumption patterns from smart meters big data", in *IEEE Electrical Power and Energy Conference 2016 - Smart Grid and Beyond : Future of the Integrated Power System*, Ottawa, ON, Canada: IEEE, Oct. 2016.

[8] K. Jack and K. William, "The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes", *Scientific Data*, vol. 2, no. 150007, 2015. DOI: 10.1038/sdata.2015.7. arXiv: 1404.0284.

[9] S. Makonin, B. Ellert, I. V. Bajic, and F. Popowich, "Ampds2 - almanac of minutely power dataset : Electricity, water, and natural gas consumption of a residential house in canada from 2012 to 2014", *Scientific Data*, vol. 3, no. 160037, pp. 1–12, 2016. DOI: DOI:10.1038/sdata.2016.37.

[10] U. S. F. E. R. C. U. S. F. E. R. Commission, "Assessment of demand response and advanced metering". [Online]. Available: http://www.ferc.gov/legal/staff-reports/12-08-demand-response.pdf.

[11] O. E. D. C. LLC, *Smart grid technology*. [Online]. Available: http://www.oncor.com/EN/Pages/Smart-Grid-Technology.aspx.

[12] O. of Electricity Delivery and E. Reliability, *Smart grid*. [Online]. Available: http://energy.gov/oe/services/technology-development/smart-grid.

[13] S. E. technology platform for the electricity networks of the future, *Publication of the report on âăijsmart grid security certification in europe: Challenges and recommendationsâăi*. [Online]. Available: http://www.smartgrids.eu/News_2014_and_before.

[14] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The kdd process for extracting useful knowledge from volumes of data", *Commun. ACM*, vol. 39, no. 11, pp. 27–34, Nov. 1996, ISSN: 0001-0782. DOI: 10.1145/240455.240464. [Online]. Available: http://doi.acm.org/10.1145/240455.240464.

[15] J. Han, J. Pei, and M. Kamber, "Data mining: Concepts and techniques (third edition)", in. Morgan Kaufmann, Jun. 2011, ch. Chapter 2: Data Preprocessing, pp. 243–278, ISBN: 9780123814791. [Online]. Available: http://www.sciencedirect.com/science/book/9780123814791.

[16] J. Han, J. Pei, and M. Kamber, "Data mining: Concepts and techniques (third edition)", in. Morgan Kaufmann, Jun. 2011, ch. Chapter 1: Introduction, pp. 243–278, ISBN: 9780123814791. [Online]. Available: http://www.sciencedirect.com/science/book/9780123814791.

[17] A. G.-V. of Analytics Zementis Inc, *Predicting the future, part 2: Predictive modeling techniques*, Jun. 2012. [Online]. Available: http://www.ibm.com/developerworks/library/ba-predictive-analytics2/.

[18] J. Han, J. Pei, and M. Kamber, "Data mining: Concepts and techniques (third edition)", in. Morgan Kaufmann, Jun. 2011, ch. Chapter 6: Mining Frequent Patterns, Associations, and Correlations: Basic Concepts and Methods, pp. 243–278, ISBN: 9780123814791. [Online]. Available: http://www.sciencedirect.com/science/book/9780123814791.

[19] E. Nazerfard and D. J. Cook, "Using bayesian networks for daily activity prediction", in *Proceedings of the 13th AAAI Conference on Plan, Activity, and Intent Recognition*, ser. AAAIWS'13-13, AAAI Press, 2013, pp. 32–38. [Online]. Available: http://dl.acm.org/citation.cfm?id=2908241.2908246.

[20] N. Tanaka, "Technology roadmap - smart grids", Tech. Rep., 2011. [Online]. Available: https://www.iea.org/publications/freepublications/publication/technology-roadmap-smart-grids.html.

[21] A. Molina-Markham, P. Shenoy, K. Fu, E. Cecchet, and D. Irwin, "Private memoirs of a smart meter", in *Proceedings of the 2Nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building*, ser. BuildSys '10, Zurich, Switzerland: ACM, 2010, pp. 61–66, ISBN: 978-1-4503-0458-0. DOI: 10.1145/1878431.1878446. [Online]. Available: http://doi.acm.org/10.1145/1878431.1878446.

[22] S. S. Intille, "The goal: Smart people, not smart homes", in *In Proceedings of the International Conference on Smart Homes and Health Telematics, IOS*, Press, 2006. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?DOI=10.1.1.70.8273.

[23] G. Wood and M. Newborough, "Dynamic energy-consumption indicators for domestic appliances: Environment, behaviour and design", *Energy and Buildings*, vol. 35, no. 8, pp. 821–841, 2003, ISSN: 0378-7788. DOI: http://dx.DOI.org/10.1016/S0378-7788(02)00241-4. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378778802002414.

[24] G. Wood and M. Newborough, "Influencing user behaviour with energy information display systems for intelligent homes", *International Journal of Energy Research*,

vol. 31, no. 1, pp. 56–78, 2007, ISSN: 1099-114X. DOI: 10.1002/er.1228. [Online]. Available: http://dx.DOI.org/10.1002/er.1228.

[25] L. Hawarah, S. Ì. Ploix, and M. Han Jacomino, "Artificial intelligence and soft computing: 10th international conference, icaisc 2010, zakopane, poland, june 13-17, 2010. part i", in. Springer-Verlag Berlin Heidelberg, Germany, Jun. 2010, vol. 6113, ch. User Behavior Prediction in Energy Consumption in Housing Using Bayesian Networks, pp. 372–379, ISBN: 9783642132087. DOI: 10.1007/978-3-642-13208-7_47. [Online]. Available: http://dx.DOI.org/10.1007/978-3-642-13208-7_47.

[26] S. Rollins and N. Banerjee, "Using rule mining to understand appliance energy consumption patterns", in *Pervasive Computing and Communications (PerCom), 2014 IEEE International Conference on*, Mar. 2014, pp. 29–37. DOI: 10.1109/PerCom.2014.6813940.

[27] C. Chelmis, J. Kolte, and V. K. Prasanna, "Big data analytics for demand response: Clustering over space and time", in *Big Data (Big Data), 2015 IEEE International Conference on*, Oct. 2015, pp. 2223–2232. DOI: 10.1109/BigData.2015.7364011.

[28] O. Ardakanian, N. Koochakzadeh, R. P. Singh, L. Golab, and S. Keshav, "Computing electricity consumption profilesfrom household smart meter data", in *Proceedings of the Workshops of the EDBT/ICDT 2014 Joint Conference (EDBT/ICDT 2014)*, Mar. 2014, pp. 140–147. [Online]. Available: http://ceur-ws.org/Vol-1133#paper-22.

[29] C. Chen and D. J. Cook, *Behavior-based home energy prediction*. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?DOI=10.1.1.420.3152.

[30] N. C. Truong, J. McInerney, L. Tran-Thanh, E. Costanza, and S. D. Ramchurn, "Forecasting multi-appliance usage for smart home energy management", in *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, ser. IJCAI '13, Beijing, China: AAAI Press, 2013, pp. 2908–2914, ISBN: 978-1-57735-633-2. [Online]. Available: http://dl.acm.org/citation.cfm?id=2540128.2540547.

[31] A. Albert and R. Rajagopal, "Smart meter driven segmentation: What your consumption says about you", *IEEE Transactions on Power Systems*, vol. 28, no. 4, pp. 4019–4030, Nov. 2013, ISSN: 0885-8950. DOI: 10.1109/TPWRS.2013.2266122.

[32] C. O. Adika and L. Wang, "Autonomous appliance scheduling based on time of use probabilities and load clustering", in *2012 10th International Power Energy Conference (IPEC)*, Dec. 2012, pp. 42–47. DOI: 10.1109/ASSCC.2012.6523236.

[33] H. S. Cho, T. Yamazaki, and M. Hahn, "Aero: Extraction of user's activities from electric power consumption data", *IEEE Transactions on Consumer Electronics*, vol. 56, no. 3, pp. 2011–2018, Aug. 2010, ISSN: 0098-3063. DOI: 10.1109/TCE.2010.5606359.

[34] J. Clement, J. Ploennigs, and K. Kabitzsch, "Smart meter: Detect and individualize adls", in *Ambient Assisted Living: 5. AAL-Kongress 2012 Berlin, Germany, January 24-25, 2012*, R. Wichert and B. Eberhardt, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 107–122, ISBN: 978-3-642-27491-6. DOI: 10.1007/978-3-642-27491-6_8. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-27491-6_8.

[35] D. Garnier-Moiroux, F. Silveira, and A. Sheth, "Towards user identification in the home from appliance usage patterns", in *Proceedings of the 2013 ACM Conference on*

Pervasive and Ubiquitous Computing Adjunct Publication, ser. UbiComp '13 Adjunct, Zurich, Switzerland: ACM, 2013, pp. 861–868, ISBN: 978-1-4503-2215-7. DOI: 10.1145/2494091.2497328. [Online]. Available: http://doi.acm.org/10.1145/2494091.2497328.

[36] J. Alcalá, O. Parson, and A. Rogers, "Detecting anomalies in activities of daily living of elderly residents via energy disaggregation and cox processes", in *Proceedings of the 2Nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments*, ser. BuildSys '15, Seoul, South Korea: ACM, 2015, pp. 225–234, ISBN: 978-1-4503-3981-0. DOI: 10.1145/2821650.2821654. [Online]. Available: http://doi.acm.org/10.1145/2821650.2821654.

[37] X. Zhang, T. Kato, and T. Matsuyama, "Learning a context-aware personal model of appliance usage patterns in smart home", in *2014 IEEE Innovative Smart Grid Technologies - Asia (ISGT ASIA)*, May 2014, pp. 73–78. DOI: 10.1109/ISGT-Asia.2014.6873767.

[38] Y. C. Chen, Y. L. Ko, and W. C. Peng, "An intelligent system for mining usage patterns from appliance data in smart home environment", in *2012 Conference on Technologies and Applications of Artificial Intelligence*, ser. TAAI '12, Washington, DC, USA: IEEE Computer Society, Nov. 2012, pp. 319–322, ISBN: 978-0-7695-4919-4. DOI: 10.1109/TAAI.2012.54. [Online]. Available: http://dx.doi.org/10.1109/TAAI.2012.54.

[39] Y.-C. Chen, Y.-L. Ko, W.-C. Peng, and W.-C. Lee, "Mining appliance usage patterns in smart home environment", in *Advances in Knowledge Discovery and Data Mining: 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14-17, 2013, Proceedings, Part I*, J. Pei, V. S. Tseng, L. Cao, H. Motoda, and G. Xu, Eds. Berlin,

Heidelberg: Springer Berlin Heidelberg, 2013, pp. 99–110, ISBN: 978-3-642-37453-1. DOI: 10.1007/978-3-642-37453-1_9. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-37453-1_9.

[40] J. Liao, L. Stankovic, and V. Stankovic, "Detecting household activity patterns from smart meter data", in *Intelligent Environments (IE), 2014 International Conference on*, Jun. 2014, pp. 71–78. DOI: 10.1109/IE.2014.18.

[41] K. Ellegard, J. WidÃl'n, and K. Vrotsou, "Appliances facilitating everyday life - electricity use derived from daily activities", in *World Renewable Energy Congress - Sweden; 8-13 May; 2011; LinkÃűping; Sweden*, May 2011, pp. 1031–1038, ISBN: 978-91-7393-070-3. DOI: 10.3384/ecp110571031.

[42] T. Thomas, C. Cashen, and S. Russ, "Leveraging smart grid technology for home health care", in *2013 IEEE International Conference on Consumer Electronics (ICCE)*, 2013, pp. 274–275. DOI: 10.1109/ICCE.2013.6486892.

[43] N. Zhang, L. F. Ochoa, and D. S. Kirschen, "Investigating the impact of demand side management on residential customers", in *Innovative Smart Grid Technologies (ISGT Europe), 2011 2nd IEEE PES International Conference and Exhibition on*, Dec. 2011, pp. 1–6. DOI: 10.1109/ISGTEurope.2011.6162699.

[44] F. D. Silva and O. Mohammed, "Demand side load control with smart meters", in *2013 IEEE Power Energy Society General Meeting*, Jul. 2013, pp. 1–5. DOI: 10.1109/PESMG.2013.6673014.

[45] S. Waczowicz, M. Reischl, V. Hagenmeyer, R. Mikut, S. Klaiber, P. Bretschneider, I. Konotop, and D. Westermann, "Demand response clustering - how do dynamic prices affect household electricity consumption?", in *PowerTech, 2015 IEEE Eindhoven*, Jun. 2015, pp. 1–6. DOI: 10.1109/PTC.2015.7232493.

[46] C. Chalmers, W. Hurst, M. Mackay, and P. Fergus, "Smart meter profiling for health applications", in *2015 International Joint Conference on Neural Networks (IJCNN)*, Jul. 2015, pp. 1–7. DOI: 10.1109/IJCNN.2015.7280836.

[47] S. Haben, C. Singleton, and P. Grindrod, "Analysis and clustering of residential customers energy behavioral demand using smart meter data", *IEEE Transactions on Smart Grid*, vol. 7, no. 1, pp. 136–144, Jan. 2016, ISSN: 1949-3053. DOI: 10.1109/TSG.2015.2409786.

[48] A. Alhamoud, P. Xu, F. Englert, A. Reinhardt, P. Scholl, D. Boehnstedt, and R. Steinmetz, "Extracting human behavior patterns from appliance-level power consumption data", in *Wireless Sensor Networks: 12th European Conference, EWSN 2015, Porto, Portugal, February 9-11, 2015. Proceedings*, T. Abdelzaher, N. Pereira, and E. Tovar, Eds. Cham: Springer International Publishing, 2015, pp. 52–67, ISBN: 978-3-319-15582-1. DOI: 10.1007/978-3-319-15582-1_4. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-15582-1_4.

[49] J. Zhao, R. Yun, B. Lasternas, H. Wang, K. Lam, A. Aziz, and V. Loftness, *Occupant behavior and schedule prediction based on office appliance energy consumption data mining*, 2013. [Online]. Available: https://www.andrew.cmu.edu/user/jzhao1/papers/CISBAT2013_Zhao%20Jie.pdf.

[50] K. Basu, V. Debusschere, and S. Bacha, "Appliance usage prediction using a time series based classification approach", in *IECON 2012 - 38th Annual Conference on IEEE Industrial Electronics Society*, Oct. 2012, pp. 1217–1222. DOI: 10.1109/IECON.2012.6388597.

[51] K. Basu, L. Hawarah, N. Arghira, H. Joumaa, and S. Ploix, "A prediction system for home appliance usage", *Energy and Buildings*, vol. 67, pp. 668–679, 2013, ISSN:

0378-7788. DOI: http://dx.DOI.org/10.1016/j.enbuild.2013.02.008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378778813000789.

[52] L. Ong, M. Bergés, and H. Y. Noh, "Exploring sequential and association rule mining for pattern-based energy demand characterization", in *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings*, ser. BuildSys'13, Roma, Italy: ACM New York NY, USA, 2013, 25:1–25:2, ISBN: 978-1-4503-2431-1. DOI: 10.1145/2528282.2528308. [Online]. Available: http://DOI.acm.org/10.1145/2528282.2528308.

[53] M. Hassani, C. Beecks, D. Tows, and T. Seidl, "Mining sequential patterns of event streams in a smart home application", in *Proceedings of the LWA 2015 Workshops: KDML, FGWM, IR, and FGDB*, Trier, Germany, Oct. 2015. DOI: 10.1109/ICMLA.2015.62. [Online]. Available: http://ceur-ws.org.

[54] Y.-C. Chen, C.-C. Chen, W.-C. Peng, and W.-C. Lee, "Mining correlation patterns among appliances in smart home environment", in *Advances in Knowledge Discovery and Data Mining: 18th Pacific-Asia Conference, PAKDD 2014, Tainan, Taiwan, May 13-16, 2014. Proceedings, Part II*, V. S. Tseng, T. B. Ho, Z.-H. Zhou, A. L. P. Chen, and H.-Y. Kao, Eds. Cham: Springer International Publishing, 2014, pp. 222–233, ISBN: 978-3-319-06605-9. DOI: 10.1007/978-3-319-06605-9_19. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-06605-9_19.

[55] Y. C. Chen, H. C. Hung, B. Y. Chiang, S. Y. Peng, and P. J. Chen, "Incrementally mining usage correlations among appliances in smart homes", in *Network-Based Information Systems (NBiS), 2015 18th International Conference on*, Sep. 2015, pp. 273–279. DOI: 10.1109/NBiS.2015.43.

[56] Y. C. Chen, W. C. Peng, and S. Y. Lee, "Mining temporal patterns in time interval-based data", *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 12, pp. 3318–3331, Dec. 2015, ISSN: 1041-4347. DOI: 10.1109/TKDE.2015.2454515.

[57] Y. S. Liao, H. Y. Liao, D. R. Liu, W. T. Fan, and H. Omar, "Intelligent power resource allocation by context-based usage mining", in *Advanced Applied Informatics (IIAI-AAI), 2015 IIAI 4th International Congress on*, Jul. 2015, pp. 546–550. DOI: 10.1109/IIAI-AAI.2015.165.

[58] S. Rahimi, A. D. C. Chan, and R. A. Goubran, "Usage monitoring of electrical devices in a smart home", in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug. 2011, pp. 5307–5310. DOI: 10.1109/IEMBS.2011.6091313.

[59] S. Rahimi, A. D. C. Chan, and R. A. Goubran, "Nonintrusive load monitoring of electrical devices in health smart homes", in *Instrumentation and Measurement Technology Conference (I2MTC), 2012 IEEE International*, May 2012, pp. 2313–2316. DOI: 10.1109/I2MTC.2012.6229453.

[60] Y. DING, J. BORGES, M. A. NEUMANN, and M. BEIGL, *Using sequence mining to understand daily activity patterns for load forecasting enhancement*, 2015. [Online]. Available: https://www.teco.edu/~michael/publication/2015_ISC2_144.pdf.

[61] F. L. Quilumba, W. J. Lee, H. Huang, D. Y. Wang, and R. L. Szabados, "Using smart meter data to improve the accuracy of intraday load forecasting considering customer behavior similarities", *IEEE Transactions on Smart Grid*, vol. 6, no. 2, pp. 911–918, Mar. 2015, ISSN: 1949-3053. DOI: 10.1109/TSG.2014.2364233.

[62] J. L. Viegas, S. M. Vieira, J. M. C. Sousa, R. Melicio, and V. M. F. Mendes, "Electricity demand profile prediction based on household characteristics", in *2015 12th International Conference on the European Energy Market (EEM)*, May 2015, pp. 1–5. DOI: 10.1109/EEM.2015.7216746.

[63] K. Gajowniczek and T. Zabkowski, "Data mining techniques for detecting household characteristics based on smart meter data", *Energies*, vol. 8, no. 7, p. 7407, 2015, ISSN: 1996-1073. DOI: 10.3390/en8077407. [Online]. Available: http://www.mdpi.com/1996-1073/8/7/7407.

[64] P. Zhang, X. Wu, X. Wang, and S. Bi, "Short-term load forecasting based on big data technologies", *CSEE Journal of Power and Energy Systems*, vol. 1, no. 3, pp. 59–67, Sep. 2015. DOI: 10.17775/CSEEJPES.2015.00036.

[65] J. Kwac, J. Flora, and R. Rajagopal, "Household energy consumption segmentation using hourly data", *IEEE Transactions on Smart Grid*, vol. 5, no. 1, pp. 420–430, 2014, ISSN: 1949-3053. DOI: 10.1109/TSG.2013.2278477.

[66] C. Flath, D. Nicolay, T. Conte, C. van Dinther, and L. Filipova-Neumann, "Cluster analysis of smart metering data", *Business & Information Systems Engineering*, vol. 4, no. 1, pp. 31–39, 2012, ISSN: 1867-0202. DOI: 10.1007/s12599-011-0201-5. [Online]. Available: http://dx.doi.org/10.1007/s12599-011-0201-5.

[67] G. Grigoras and F. Scarlatache, "Processing of smart meters data for peak load estimation of consumers", in *2015 9th International Symposium on Advanced Topics in Electrical Engineering (ATEE)*, May 2015, pp. 864–867. DOI: 10.1109/ATEE.2015.7133922.

[68] S. Park, S. Ryu, Y. Choi, and H. Kim, "A framework for baseline load estimation in demand response: Data mining approach", in *Smart Grid Communications (SmartGridComm), 2014 IEEE International Conference on*, Nov. 2014, pp. 638–643. DOI: 10.1109/SmartGridComm.2014.7007719.

[69] H. N. Zala and A. R. Abhyankar, "A novel approach to design time of use tariff using load profiling and decomposition", in *Power Electronics, Drives and Energy Systems (PEDES), 2014 IEEE International Conference on*, Dec. 2014, pp. 1–6. DOI: 10.1109/PEDES.2014.7042027.

[70] M. Amin-Naseri and A. Soroush, "Combined use of unsupervised and supervised learning for daily peak load forecasting", *Energy Conversion and Management*, vol. 49, no. 6, pp. 1302 –1308, 2008, ISSN: 0196-8904. DOI: http://dx.doi.org/10.1016/j.enconman.2008.01.016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0196890408000174.

[71] S. Fan, L. Chen, and W.-J. Lee, "Machine learning based switching model for electricity load forecasting", *Energy Conversion and Management*, vol. 49, no. 6, pp. 1331 –1344, 2008, ISSN: 0196-8904. DOI: http://dx.doi.org/10.1016/j.enconman.2008.01.008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0196890408000216.

[72] H.-F. Shi and Y.-X. Lu, "Bayesian neural networks for short term load forecasting", in *2009 International Conference on Wavelet Analysis and Pattern Recognition*, Jul. 2009, pp. 160–165. DOI: 10.1109/ICWAPR.2009.5207407.

[73] P. Lauret, E. Fock, R. N. Randrianarivony, and J.-F. Manicom-Ramsamy, "Bayesian neural network approach to short time load forecasting", *Energy Conversion and Management*, vol. 49, no. 5, pp. 1156 –1166, 2008, ISSN: 0196-8904. DOI: http://dx.

doi.org/10.1016/j.enconman.2007.09.009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0196890407003032.

[74] P. Mandal, T. Senjyu, K. Uezato, and T. Funabashi, "Several-hours-ahead electricity price and load forecasting using neural networks", in *IEEE Power Engineering Society General Meeting, 2005*, Jun. 2005, 2146–2153 Vol. 3. DOI: 10.1109/PES.2005.1489530.

[75] Y. c. Guo, D. x. Niu, and Y. x. Chen, "Support vector machine model in electricity load forecasting", in *2006 International Conference on Machine Learning and Cybernetics*, Aug. 2006, pp. 2892–2896. DOI: 10.1109/ICMLC.2006.259076.

[76] W.-C. Hong, "Electric load forecasting by support vector model", *Applied Mathematical Modelling*, vol. 33, no. 5, pp. 2444 –2454, 2009, ISSN: 0307-904X. DOI: http://dx.doi.org/10.1016/j.apm.2008.07.010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0307904X08001844.

[77] T. K. Wijaya, M. Vasirani, S. Humeau, and K. Aberer, "Cluster-based aggregate forecasting for residential electricity demand using smart meter data", in *Big Data (Big Data), 2015 IEEE International Conference on*, Oct. 2015, pp. 879–887. DOI: 10.1109/BigData.2015.7363836.

[78] A. Reinhardt, D. Christin, and S. S. Kanhere, "Predicting the power consumption of electric appliances through time series pattern matching", in *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings*, ser. BuildSys'13, Roma, Italy: ACM, 2013, 30:1–30:2, ISBN: 978-1-4503-2431-1. DOI: 10.1145/2528282.2528315. [Online]. Available: http://doi.acm.org/10.1145/2528282.2528315.

[79] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation", in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, Dallas, Texas, USA*, ser. SIGMOD '00, New York, NY, USA: ACM, 2000, pp. 1–12, ISBN: 1-58113-217-4. DOI: 10.1145/342009.335372. [Online]. Available: http://DOI.acm.org/10.1145/342009.335372.

[80] J. Han, J. Pei, Y. Yin, and R. Mao, "Mining frequent patterns without candidate generation: A frequent-pattern tree approach", *Data Mining and Knowledge Discovery*, vol. 8, no. 1, pp. 53–87, 2004, ISSN: 1573-756X. DOI: 10.1023/B:DAMI.0000005258.31418.83. [Online]. Available: http://dx.DOI.org/10.1023/B:DAMI.0000005258.31418.83.

[81] J. MacQueen, "Some methods for classification and analysis of multivariate observations", in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, VOLUME 1: Statistics*, Berkeley, California: University of California Press, 1967, pp. 281–297. [Online]. Available: http://projecteuclid.org/euclid.bsmsp/1200512992.

[82] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases", in *Proceedings of the 20th International Conference on Very Large Data Bases*, ser. VLDB '94, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1994, pp. 487–499, ISBN: 1-55860-153-8. [Online]. Available: http://dl.acm.org/citation.cfm?id=645920.672836.

[83] P.-N. Tan, M. Steinbach, and V. Kumar, "Introduction to data mining", in. Pearson, May 2005, vol. 1, ch. Chapter 8 Cluster Analysis: Basic Concepts and Algorithms, pp. 487–568, ISBN: 978-0321321367.

[84] J. Han, J. Pei, and M. Kamber, "Data mining: Concepts and techniques (third edition)", in. Morgan Kaufmann, Jun. 2011, ch. Chapter 10: Cluster Analysis: Basic Concepts and Methods, pp. 443–494, ISBN: 9780123814791. [Online]. Available: http://www.sciencedirect.com/science/book/9780123814791.

[85] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis", *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987, ISSN: 0377-0427. DOI: http://dx.DOI.org/10.1016/0377-0427(87)90125-7. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0377042787901257.

[86] M. B. Al-Zoubi and M. Al Rawi, "An efficient approach for computing silhouette coefficients", *Journal of Computer Science*, vol. 4, no. 3, pp. 252–255, 2008, ISSN: 1549-3636. DOI: 10.3844/jcssp.2008.252.255. [Online]. Available: http://thescipub.com/html/10.3844/jcssp.2008.252.255.

[87] W. Haizhou and S. Mingzhou, "Ckmeans.1d.dp: Optimal k-means clustering in one dimension by dynamic programming", *The R Journal*, vol. 3, no. 2, pp. 29–33, Dec. 2011, ISSN: 1949-3053. DOI: 10.1109/TSG.2013.2278477. [Online]. Available: http://journal.r-project.org/archive/2011-2/RJournal_2011-2_Wang+Song.pdf.

[88] I. Ben-Gal, "Bayesian networks", in *Encyclopedia of Statistics in Quality and Reliability*. John Wiley & Sons, Ltd, 2008, ISBN: 9780470061572. DOI: 10.1002/9780470061572.eqr089. [Online]. Available: http://dx.DOI.org/10.1002/9780470061572.eqr089.

[89] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1988, ISBN: 0-934613-73-7.

[90] P. C. Pendharkar, G. H. Subramanian, and J. A. Rodger, "A probabilistic model for predicting software development effort", *IEEE Transactions on Software Engineering*, vol. 31, no. 7, pp. 615–624, Jul. 2005, ISSN: 0098-5589. DOI: 10.1109/TSE.2005.75.

[91] D. Heckerman, "Bayesian networks for data mining", *Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 79–119, 1997, ISSN: 1573-756X. DOI: 10.1023/A:1009730122752. [Online]. Available: http://dx.DOI.org/10.1023/A:1009730122752.

[92] D. Heckerman, "Learning in graphical models", in, M. I. Jordan, Ed., Cambridge, MA, USA: MIT Press, 1999, ch. A Tutorial on Learning with Bayesian Networks, pp. 301–354, ISBN: 0-262-60032-3. [Online]. Available: http://dl.acm.org/citation.cfm?id=308574.308676.

[93] D. Barber, *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2012, ch. Chapter 3 Belief Networks, pp. 29–57, ISBN: 978-0-521-51814-7.

[94] J. Han, J. Pei, and M. Kamber, "Data mining: Concepts and techniques (third edition)", in. Morgan Kaufmann, Jun. 2011, ch. Chapter 8: Classification: Basic Concepts, pp. 243–278, ISBN: 9780123814791. [Online]. Available: http://www.sciencedirect.com/science/book/9780123814791.

[95] J. Han, J. Pei, and M. Kamber, "Data mining: Concepts and techniques (third edition)", in. Morgan Kaufmann, Jun. 2011, ch. Chapter 9: Classification: Advanced

Methods, pp. 243–278, ISBN: 9780123814791. [Online]. Available: http://www.sciencedirect.com/science/book/9780123814791.

[96] G. Tsoumakas and I. Katakis, "Data warehousing and mining: Concepts, methodologies, tools, and applications", in. IGI Global, Hershey, PA, Jun. 2008, ch. Chapter 6: Multi-Label Classification: An Overview, pp. 64–74, ISBN: 9781599049519. DOI: 10.4018/978-1-59904-951-9.ch006. [Online]. Available: http://www.sciencedirect.com/science/book/9780123814791.

[97] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Mining multi-label data", in *Data Mining and Knowledge Discovery Handbook*, O. Maimon and L. Rokach, Eds. Boston, MA: Springer US, 2010, pp. 667–685, ISBN: 978-0-387-09823-4. DOI: 10.1007/978-0-387-09823-4_34. [Online]. Available: http://dx.DOI.org/10.1007/978-0-387-09823-4_34.

[98] E. Gibaja and S. Ventura, "A tutorial on multilabel learning", *ACM Comput. Surv.*, vol. 47, no. 3, 52:1–52:38, Apr. 2015, ISSN: 0360-0300. DOI: 10.1145/2716262. [Online]. Available: http://doi.acm.org/10.1145/2716262.

[99] M. S. Sorower, "A literature survey on algorithms for multi-label learning", Tech. Rep., 2010. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.364.5612.

[100] J. Read, "Multi-label classification", 2015, [Online]. Available: http://jmread.github.io/talks/Tutorial-MLC-Porto.pdf.