Received July 24, 2015, accepted August 12, 2015, date of publication August 24, 2015, date of current version September 1, 2015.

Digital Object Identifier 10.1109/ACCESS.2015.2472355

Rule Induction-Based Knowledge Discovery for Energy Efficiency

QIPENG CHEN[1], (Student Member, IEEE), ZHONG FAN[2], DRITAN KALESHI[3], AND SIMON ARMOUR[1]

[1]Communication Systems and Networks Group, Department of Electrical and Electronic Engineering, University of Bristol, Bristol BS8 1TH, U.K.
[2]Telecommunications Research Laboratory, Toshiba Research Europe Ltd., Bristol BS1 4ND, U.K.
[3]Digital Catapult, London NW1 2RA, U.K.

Corresponding author: Q. Chen (qipeng.chen@bristol.ac.uk)

This work was supported in part by the U.K. Engineering and Physical Sciences Research Council (EPSRC) HubNet Project under Grant EP/I013636/1, in part by EPSRC through Dorothy Hodgkin Scholarship, and in part by Toshiba Research Europe Ltd.

ABSTRACT Rule induction is a practical approach to knowledge discovery. Provided that a problem is developed, rule induction is able to return the knowledge that addresses the goal of this problem as if-then rules. The primary goals of knowledge discovery are for prediction and description. The rule format knowledge representation is easily understandable so as to enable users to make decisions. This paper presents the potential of rule induction for energy efficiency. In particular, three rule induction techniques are applied to derive knowledge from a dataset of thousands of Irish electricity customers' time-series power consumption records, socio-demographic details, and other information, in order to address the following four problems: 1) discovering mathematically interesting knowledge that could be found useful; 2) estimating power consumption features for customers, so that personalized tariffs can be assigned; 3) targeting a subgroup of customers with high potential for peak demand shifting; and 4) identifying customer attitudes that dominate energy conservation.

INDEX TERMS Energy efficiency, knowledge discovery, smart grids, subgroup discovery.

I. INTRODUCTION

Knowledge Discovery in Databases (KDD) intends to discover potential knowledge from data [1]. A KDD problem is commonly driven by a Goal with respect to a Target, where the goal is for either prediction or description, and the target is a user-interested variable. For example, for a dataset of customers' socio-demographic and electric power consumption records, individual customer's overall power consumption can be taken as the target. Subsequently, a predictive goal could be discovering a model that guides the classification of low, medium and high power consumers based on their socio-demographic characteristics; a descriptive goal could be identifying customers in terms of characteristics which indicate the potential to save energy. However, a KDD problem sometimes has no target (e.g., the intention is to discover knowledge that is potentially useful rather than for a specific target), for which case the goal can only be descriptive, and the knowledge that is ''mathematically interesting'' will be extracted and presented to domain experts for them to determine usefulness. Either case needs an intuitive representation of knowledge, so users can understand and use the knowledge effectively. Rule induction returns knowledge as if-then rules that are natural and easily understandable, so it is considered to be a practical approach for knowledge discovery.

Association Rule Learning and Classification Rule Learning are two typical rule induction techniques, where the former learns from data without target and produces descriptive rules about variable regularities [2] and the latter learns from data with a target and produces predictive rules with respect to the target [3]. Subgroup discovery is another rule induction technique which lies halfway between association and classification rule learning. It learns from data with target but generates descriptive rules [4]. In this paper, we show the potential of these three rule induction techniques for energy suppliers to provide peak electric power demand shifting, electricity energy conservation and other energy efficiency related services.

Domestic households form a target group which constitutes a big portion of the overall electric power consumption, e.g., nearly 30% of UK electricity is consumed by domestic households [5]. In recent years, various attempts for household energy efficiency have been proposed, which include employing Time-of-Use (TOU) tariffs, implementing

Demand Side Management (DSM) and sending feedback to customers. All these attempts are made available by Smart Grid advances, such as the introduction of smart meters and the incorporated data acquisition and communication infrastructures.

Energy suppliers can tell customers when it is cheap to use electricity. Commonly, higher electricity prices can be assigned to peak times, and lower prices are assigned to off-peak times under what are referred to as TOU tariffs [6]. Consequently, most customers may shift power demand out from peak times and consume more power at off-peak times, which may lead to new peaks. This has been proved by the simulation result in [7], so one can assign customers different tariffs to avoid such a situation, e.g., assigning the customers of higher night time power consumption a lower price for the daytime, and assigning customers of stable daily consumption a flat tariff. Chicco *et al.* [8] first classified all customers into different categories according to their historical power consumption records, then assign dedicated tariffs to different categories of customers. It can be seen that assigning a customer a tariff requires the information about the features of this customer's power consumption patterns, but such information is only available for the customers with smart meters. Nevertheless, for the customers without smart meters, certain power consumption features can still be estimated. For example, in [9], four power consumption features (i.e., total power consumption, the maximum half-hourly consumption, the averaged daily load factor and the frequently appeared peak time) are estimated on the basis of customers' dwelling and occupant characteristics.

There is also an increasing interest in applying DSM for shifting peak power demand [10]. DSM enables automatic appliance control which switches off appliances at peak times and switches on appliances during off-peak times. The appliance control also takes into account customers' feelings of comfort thereby avoiding violating customers' routine living style.

Additionally, sending customers feedback can also improve energy efficiency, especially for energy conservation. And both the feedback contents and the way of provision determine the effectiveness. Basic feedback includes bills, but enhanced information like a bill comparison with the past [11] or peers [12], a summary of recent appliance usage activities [13] and energy tips [14] can more likely draw customers' attention to energy issues. Instead of generalised energy tips for all customers, energy tips should be personalized according to customers' characteristics and power consumption features [15]. For example, in [14] customer statistics with respect to various characteristics (i.e., households' appliance and demographic characteristics, and occupants' awareness of energy issues) are studied, and potential for energy conservation are pointed out (e.g., the proportion of customers that do not know their appliances' energy ratings is 84%. For these customers, their awareness of appliance ratings could be increased via feedback provision). Customers' preferred method of feedback provision is studied in [16] where the preference is found to be dependent on customers' ages - for example, older people prefer to receiving feedback by letter rather than in-home display and email. It is also found that older people use more electricity, and the authors explain it is probably because older people spend more time at home. In [17], how each characteristic is correlated with the power consumption is analysed, in which the potential of the change of customers' attitudes for energy conservation is claimed. Therefore, energy suppliers can consider guiding customers to change attitudes via feedback provision, so as to save energy. As discussed, multiple characteristics affect power consumptions indirectly, which should be taken into account when sending feedback. In case certain characteristics of a customer are unknown, using power consumption features to predict characteristics is proved to be possible in [18].

The following scientific questions arise in accordance with the above survey.

1) *Question One:* As described in [16], the relation between customer age and the amount of time spend at home explains why the consumption feature (i.e., high power consumption) has occurred. The question is, how can we discover more such relations that are potentially useful?
2) *Question Two:* In [9] customer consumption features are estimated based on their dwelling and occupant characteristics. The question is: can the mapping between characteristics and consumption features be represented more intuitively, like a manual? So mathematical calculations such as those in [9] are not required when estimating a new customer's consumption features and assigning this customer a tariff.
3) *Question Three:* Implementing the DSM for every single household will be a gradual procedure due to high costs. The question is: can we target a subgroup of households for which DSM can make significant energy efficiency improvements? So energy suppliers can consider installing DSM for these customers first.
4) *Question Four:* In [17] correlations between energy conservation and individual attitudes are analysed separately, while the energy conservation is actually affected by multiple cooperative attitudes. The question is: can we identify the set of attitudes that dominate energy conservation, so energy suppliers can know how to change customers' attitudes toward saving energy.

An Electricity Customer Behavior Trial (ECBT) dataset is made available to the public by the Irish Commission for Energy Regulation (CER), which records thousands of customers' power consumptions along with their dwelling, occupant, appliance and attitude characteristics [19]. Such data hides a vast amount of knowledge, which potentially includes that which can answer the above questions. We thereby formulate each question as a KDD problem.

Provided a problem is given and the nature of this problem is determined (i.e., if the goal is predictive or descriptive, if a target is chosen), an appropriate rule induction technique can

be chosen accordingly. The optimal knowledge that addresses the problem will be automatically discovered from the ECBT data and will be presented as if-then rules. This shows the benefits of rule induction techniques compared to the other methods. For example, rule induction techniques are able to identify the complete set of attitudes that dominate energy conservation in one step. However, if we use the method in [17], we need to first analyse the correlations between each individual attitude and power consumption separately and then manually determine which characteristics should be in the set and which ones should be removed.

The four formulated KDD problems have covered all three types of knowledge discovery cases, namely the case with a descriptive goal and no target, the case with a predictive goal and a target and the case with a descriptive goal and a target. Association rule learning is used for the first problem with the goal of discovering descriptive rules about variable regularities. Classification rule learning is used for the second problem, where the goal is predictive and the target could be a power consumption feature. Subgroup discovery is for the third problem, where the goal is to describe customers in terms of what kinds of characteristics will mostly benefit from DSM, and the target is a variable that denotes how much the customer can benefit from DSM. The fourth problem is solved by subgroup discovery as well, where the goal is to describe customers in terms of what kinds of attitudes perform better on energy conservation, and the target is a variable denoting customers' energy conservation capabilities. In this study, we use the Apriori algorithm [20], Decision Tree (ID3) [21] and CN2-MSD [22] to discover association, classification and subgroup discovery rules respectively.

The contributions of this paper are as follows:

- To the best of our knowledge, this is the first comprehensive study on using rule induction techniques for energy efficiency;
- This study provides guidelines on how to apply a rule induction technique to solve a specific problem in power system engineering;

The rest of this paper is organised as follows. The ECBT dataset along with the data pre-processing are introduced in Section II. The problem solving processes and the applied rule induction techniques and algorithms are introduced in Section III. The experimental results are summarized in Section IV. Conclusions are drawn in Section V.

II. ELECTRICITY CUSTOMER BEHAVIOR DATASET

We first introduce the whole dataset in this section. The exact usage of data for each specific problem will be recalled in the next section when that problem is described.

Between 2009 and 2010, the Irish CER conducted electricity customer behavior trials with 4225 residential and 2220 other customers [19]. Smart meters were installed at these customers' premises, which measure half-hourly power consumptions. The smart meter data were collected since 14-07-2009, and corresponding services including TOU tariffs and demand response stimuli were provided since 01-01-2010. Before the trials, there was a survey which gathered all the residential customers' dwelling, occupant, appliance and attitude characteristics. In this study, we focus on residential customers, which yields 3488 instances. We use the consumption data of weekdays between 22-09-2009 and 20-12-2009 (i.e. the autumn period) for the study. In the following we show the selection of power consumption features and customer characteristics.

A. POWER CONSUMPTION FEATURES

A customer's power consumption features can be interpreted by certain metrics. Six metrics have been studied, where the first four are from [9] that are primarily for load forecasting, and the last two are proposed in [23] for load profiling.

1) *Total Power Consumption:* It is a customer's total power consumption that is given by Equation (1), where E_i^j denotes the i^{th} half-hourly power consumption measurement of day j, m is the number of days and n is the number of power consumption measurements per day. In our case, m and n are equal to 64 and 48 respectively.

$$\text{TotalkWh} = \sum_{j=1}^{m} \sum_{i=1}^{n} E_i^j. \tag{1}$$

2) *Averaged Daily Maximum Demand:* This is the average of multiple days' maximum half-hourly loads, which is given by Equation (2), where $E_{i_{\max}}^j$ is the maximum half-hourly power consumption of day j, and $i_{\max}$ is the half-hourly slot of the day that $E_{i_{\max}}^j$ happens. Multiplying $E_{i_{\max}}^j$ with 2 denotes the average load of the interval.

$$E_{i_{\max}}^j = \max\{E_i^j, i \in [1, n]\},$$
$$\text{ADMD} = \frac{1}{m} \sum_{j=1}^{m} 2E_{i_{\max}}^j. \tag{2}$$

3) *Averaged Daily Load Factor:* Load factor indicates the stability of power consumption. The average of multiple days' load factor is given by Equation (3). The maximum load factor is 1, which means power consumptions over time are even. The smaller the load factor, the more dynamic.

$$\text{ADLFactor} = \frac{1}{m} \sum_{j=1}^{m} \frac{\sum_{i=1}^{n} E_i^j}{2E_{i_{\max}}^j \times 24}. \tag{3}$$

Please note that the ADMD and ADLFactor in [9] are slightly different from the definitions of Maximum demand and Load factor in power system engineering. The standard maximum demand is the maximum measured load of a whole period rather than a day and so is the load factor. By using ADMD and ADLFactor, the impact of extreme measurements can be avoided.

4) *Time of Use:* It is the most likely time of the day that the maximum half-hourly power consumption happens. It helps to determine if a customer is a peak time electricity user. The mathematical definition of TOU

is given by Equation (4), where the mode operation produces the element that appears most often in a set.

$$\text{TOU} = \text{mode}\{i_{\max} \mid E^j_{i_{\max}}, j \in [1, m]\}. \quad (4)$$

5) *Lunch Impact:* It is the ratio of the power consumptions between the lunch time (i.e., 12:00 to 14:00) and the full day. The average Lunch Impact of m days is given by Equation (5).

$$\text{LunImpact} = \frac{1}{m}\sum_{j=1}^{m}\frac{\frac{1}{4}\sum_{i=25}^{28}E_i^j}{\frac{1}{n}\sum_{i=1}^{n}E_i^j}. \quad (5)$$

6) *Evening Impact:* This highlights the proportion of power consumed in the evening (i.e., 18:00 to 22:00). The mathematical definition of Evening Impact is given by Equation (6).

$$\text{EveImpact} = \frac{1}{m}\sum_{j=1}^{m}\frac{\frac{1}{8}\sum_{i=37}^{44}E_i^j}{\frac{1}{n}\sum_{i=1}^{n}E_i^j}. \quad (6)$$

Finally, we discretize all these power consumption features, because they are numerical type variables which are not applicable to rule induction problems. We convert TOU to a binary variable (1 & 0) to distinguish the peak (07:30-09:00 and 17:30-20:00) and off-peak times. For each of the other features, it is discretized to Low, Medium and High levels based on the magnitudes of feature values, where the three levels are denoted by 1, 2 and 3 respectively. For example, the original value of TotalkWh is between 2000 and 8000 kWh. The TotalkWh may be discretized to [2000, 3000), [3000, 6000) and [6000, 8000] levels, and an example value 5000 kWh belongs to the Medium level. We use Self-Organized Map (SOM) to discretize features. SOM is an unsupervised network that learns the topology and distribution of the training input and returns an optimal strategy for clustering [24]. Using SOM for numerical variable discretization is proposed in [25].

Therefore, each feature can actually categorise customers into a certain number of groups. In power system engineering practice, the customers are most commonly grouped based on their time-series power consumption curves. We have also done such a grouping, and the procedure is described in III-B-(2).

B. CHARACTERISTICS

For every residential customer the following information has been gathered by CER prior to the trials: the customer's dwelling characteristics, the occupant characteristics, the appliances' situations, the occupants' attitudes and awareness of energy issues.

1) DWELLING AND OCCUPANT CHARACTERISTICS

In this part, we describe the first three types of characteristics (hereafter referred to as dwelling and occupant characteristics). We try to retain as many characteristic records as possible, but a few characteristics are still removed because: firstly, data are missing for some characteristics, e.g., the *dwelling floor area* is removed; secondly, a few characteristics' values are extremely imbalanced, e.g., the positive answers to the survey question *if the attic is insulated* account for more than 90% of all customers' responses; thirdly, multiple characteristics show redundant information, e.g., *if there is Internet in the dwelling* is actually covered by *is the Internet used regularly*, so it is removed. Furthermore, in the original record the number and the usage frequency of each individual appliance (e.g., washing machine, TV) are described. Instead, we build characteristics to denote the total number and using frequencies for all of the home-appliances rather than these details of any individual appliance, because this study focuses on the indirect relations between customer characteristics and power consumptions.

We show the selected characteristics in Table 1. Characteristics' names are in the first column, and brief explanations are in the second one. In the third column we show the discrete characteristic values. For numerical characteristics, those shown in column three are their discretized values. The fourth column gives the meanings of these values. And we also show the actual value ranges for numerical characteristics in the last column along with their units. We provide both the actual numerical and the discretized values for numerical dwelling and occupant characteristics, because different problems have different data requirements (i.e., classification rule learning handles numerical variables while association rule learning cannot).

2) ATTITUDE CHARACTERISTICS

In addition to the dwelling and occupant characteristics above, the ECBT data also records 31 characteristics that describe customers' subjective feelings (hereafter referred to as customer attitudes). As described by the instruction report of this data, these attitude characteristics belong to six categories:

- category one includes three characteristics about customers' attitudes toward energy and bills;
- category two includes five characteristics about customers' thought of their efforts for energy conservation;
- category three includes seven characteristics about customers' feelings of challenges for energy conservation;
- category four includes four characteristics about customers' expectations about the participation in the trial;
- category five includes six characteristics about customers' thought of the consequences of their participation;
- category six includes six characteristics about customers' satisfactions of the current electricity services.

These attitude characteristics' values are described as follows. Category four characteristics have binary values. Most of the other characteristics are on a scale of 1 to 5 where 1 is very satisfied (or strongly agree) and 5 is very dissatisfied

TABLE 1. Dwelling and occupant characteristics.

Variable Name	Variable Description	Discrete	Value Description	Numerical
voice_sex	sex from the voice	{1,2}	{male; female}	
empl_stat	employment status of chief income earner	{0,1}	{unemployed; employed}	
soc_class	social class of chief income earner	{1,2,3,4}	{A & B; C1 & C2; D & E; F}	
voice_net	Do you use internet regularly?	{0,1}	{no; yes}	
occ_net	Does any occupant use internet regularly?	{0,1}	{no; yes}	
dwg_type	Does the dwelling connect to others?	{0,1}	{no; yes}	
dwg_own	Is the dwelling owned without mortgage?	{0,1}	{no; yes}	
dwg_heat	source of home heating	{0,1}	{other; electric}	
heater_timer	Is there a timer for home heating?	{0,1}	{no; yes}	
water_heat	source of water heating	{0,1}	{other; electric}	
water_timer	Is there a timer for water heating?	{0,1}	{no; yes}	
immersion	Do you use immersion when heating is off?	{0,1}	{no; yes}	
elec_cook	Is electric used for cooking?	{0,1}	{no; yes}	
warm	Is the home kept warm?	{0,1}	{no; yes}	
dwg_age	age of your dwelling	{1,2}	{no more than 30; more than 30}	1-813 years
voice_age	age from the voice	{1,2,3}	{18-35; 36-65; 65+}	1-6 age band
occ	number of occupants / describe occupants	{1,2,3}	{single; adults (15+); adults and children}	1-12 people
occ+	number of 15+ occupants	{1,2,3}	{one; two; three & more}	1-7 people
occ-	number of 15- occupants	{0,1,2}	{none; one; two & more}	0-6 people
home_occ	number of occupants normally at home	{0,1,2,3}	{none; one; two; three & more}	0-12 people
home_occ+	number of 15+ occupants normally at home	{0,1,2}	{none; one; two & more}	0-7 people
home_occ-	number of 15- occupants normally at home	{0,1}	{none; one & more}	0-6 people
bedrooms	number of bedrooms	{1,2,3,4}	{one & two; three; four; five & more}	1-5 bedroom
appl	number of appliances in your home	{1,2,3}	{no more than 5, in between, no less than 8}	0-19 appliances
appl_use	How often are the appliances used?	{1,2}	{no more than 9 times; no less than 10 times}	0-28 times/day
ent_appl	number of entertainment appliances	{1,2,3}	{no more than 2; in between; no less than 5}	0-18 appliances
ent_use	How often are the appliances used?	{1,2,3}	{no more than 4; in between; no less than 9}	0-20 times/day
light_bulbs	proportion of energy reduction light bulbs	{1,2}	{up to half; three quarters & more}	1-5 portion
education	education level of chief income earner	{1,2,3}	{low; medium; high}	1-5 education level
income	family income level	{1,2}	{low; high}	1-5 income band

(or strongly disagree). Four attitudes' values do not follow the above provision:

- A characteristic from category two has binary values (i.e., if the customer thinks his/her previous efforts reduced bills).
- A characteristic from category three is on a scale of 1 to 6 that denote 0%, less than 5%, 5%-10%, 10%-20%, 20%-30%, and more than 30% respectively (i.e how much does the customer believe he/she could reduce electricity by).
- A characteristic from category five is on a scale of 1 to 3 that denote no change, increase, and decrease respectively (i.e., how does the customer think the bills will change as part of the trial).
- A characteristic from category five is on a scale of 0 to 6 that denote increase, no change, decrease less than 5%, decrease 5%-10%, decrease 10%-20%, decrease 20%-30% and decrease more than 30% respectively (i.e., by what amount does the customer think the bills will change as part of the trial).

We remove one characteristic (i.e., if the customer thinks he/she can learn how to reduce their bill) from category four, because more than 90% of customers' responses to this characteristic are positive which is extremely imbalanced.

C. DATA REPRESENTATION

In the machine learning domain, data is stored as a matrix. The columns are Attributes, and rows are Instances. We take the characteristics in II-B and power consumption features in II-A as examples to explain how data is built for association and classification rule learning and subgroup discovery.

For an association rule learning problem, all 36 variables (i.e., 30 characteristics and 6 features) could be taken as attributes. Each customer corresponds to an instance recording the customer's attributes' values. An instance j is interpreted as: $\{\Phi^j_1, \ldots, \Phi^j_{i_1}, \ldots, \Phi^j_{30}, \Gamma^j_1, \ldots, \Gamma^j_{i_2}, \ldots, \Gamma^j_6\}$, where $\Phi^j_{i_1}$ and $\Gamma^j_{i_2}$ are the values of the i^{th}_1 dwelling and occupant characteristic and the i^{th}_2 power consumption feature respectively. The full data is the set of all instances.

A classification rule learning problem or a subgroup discovery problem needs a target to be defined. Say the target of a classification problem is TOU, then the problem is to classify the customers as peak or off-peak time electricity users on the basis of their dwelling and occupant characteristics. The TOU values (0 & 1) are called Target Labels. The j^{th} instance is represented as $\{\Phi^j_1, \ldots, \Phi^j_{i_1}, \ldots, \Phi^j_{30}, \Gamma^j_{i_2}\}$ with 30 attribute values and the target label $\Gamma^j_{i_2}$.

III. RULE INDUCTION METHODS AND PROBLEMS

Rule induction interprets knowledge as *head*←*body* format. The head is commonly an attribute/value pair (e.g., $X = a$ or $Y \leq b$, where X and Y are attributes, and a and b are their values). The body can either be another pair or a conjunction of multiple pairs (i.e., $X = a \wedge Y \leq b$). We call an attribute/value pair and a conjunction of multiple

pairs as a Sector and a Complex respectively. In each of the following subsections, we describe the principle, heuristic and corresponding algorithm for association rule learning, classification rule learning and subgroup discovery respectively. The use of these techniques for those problems discussed in Section I are introduced.

A. ASSOCIATION RULE LEARNING

Association rule learning discovers frequent complexes. For example, complex TotalkWh=1 $\wedge$ ADMD=1 appears in 30% of customers. Provided a complex is found, multiple rules can be built based on this complex: each selector is taken as the rule head in turn, and the remaining selectors are taken as the rule body. All subsequent rules' heuristic scores are checked against a minimum threshold score. Those rules with higher scores are finally provided to the users.

An association rule satisfies two criteria: the rule's frequency should be bigger than the minimum frequency, and the rule's heuristic score should be bigger than the minimum score. A rule's frequency is commonly reflected by the metric Support which denotes the ratio of the number of the instances that are covered by this rule to the population size. The score of the heuristic is measured by user's interests. Two commonly applied heuristics are Confidence and Lift. Confidence of a rule is given by Equation (7),

$$\text{conf}(\text{head} \leftarrow \text{body}) = \frac{\text{supp}(\text{head} \cup \text{body})}{\text{supp}(\text{body})}, \tag{7}$$

where supp(head $\cup$ body) and supp(body) are the supports of this rule and this rule's body respectively. A high confidence indicates that the occurrence of the body can likely imply the occurrence of the head. However, it does not necessarily mean the head is caused by the body. Lift places more emphasis on the dependency of the head on the body, which is given by Equation (8).

$$\text{lift}(\text{head} \leftarrow \text{body}) = \frac{\text{supp}(\text{head} \cup \text{body})}{\text{supp}(\text{head}) \times \text{supp}(\text{body})}. \tag{8}$$

1) APRIORI ALGORITHM

This algorithm is initially proposed for the problem of Market Basket Transactions [20], which intends to identify sets of items that are usually purchased together. We thus use a transaction problem as an example to describe the procedure of the algorithm. We call a set of items as an itemset, and an itemset of bigger frequency than the minimum support as a frequent itemset.

Let us say the data is about the transaction records of Milk, Bread and Egg, based on which three size one itemsets (1-itemsets): {Milk}, {Bread}, {Egg} are built. And we assume all the 1-itemsets are frequent itemsets. Apriori then constructs three 2-itemsets based on the frequent 1-itemsets, which are {Milk, Bread}, {Bread, Egg} and {Milk, Egg} respectively. The three new itemsets are called candidate 2-itemsets. Apriori then removes all the candidate 2-itemsets which are not frequent. And only those remaining ones are taken as the frequent 2-itemsets and are used for constructing candidate 3-itemsets. Apriori iteratively constructs candidate k-itemsets on the basis of the frequent (k-1)-itemsets, and then finds the frequent k-itemsets from the candidate k-itemsets (k denotes the number of items in a set) until all the frequent itemsets are found. After Apriori, all the possible rules for each frequent itemset are built and checked against the minimum heuristic score. Only the rules with higher scores than the minimum score are retained.

2) PROBLEM ONE

For the first scientific question discussed in Section I, we consider if there are more rules among characteristics that are potentially useful. It is also sensible to assume there are such rules among power consumption features, and even between characteristics and features.

Therefore, we apply the Apriori algorithm to the data that takes customers' dwelling and occupant characteristics and also their power consumption features as attributes, in order to discover as many association rules as possible. The discovered rules can be provided to energy suppliers to determine the usefulness. Detailed data specifications for this problem and for the rest three problems are summarised in III-D and the results of this problem are shown in IV-A.

B. CLASSIFICATION RULE LEARNING

Classification rule learning learns from instances with labels and generates *Class*$\leftarrow$*Condition* rules. Condition is a complex, and class is a specific target label. The rule thus implies a mapping from the complex to the target label. Therefore, the label of a new instance can be determined on the basis of classification rules.

1) DECISION TREE (ID3)

Given a data with a target, ID3 builds a tree-like graph which interprets all possible mappings from different complexes to target labels. The tree is composed by multiple branches, where each branch is actually a graphical representation of a classification rule. Therefore, the tree itself is a set of classification rules.

A grown tree is connected by internal nodes, edges and leaves (terminal nodes). It starts from an internal node that is called the root. Multiple edges are then extended from the root. And each edge ends as either a leaf or the root of a new tree. The new tree then grows in the same way. The decision tree keeps growing until all the edges are terminated at leaves. An internal node of the tree is an attribute, and the edges from the node are the values of this attribute. Therefore, a node and an edge from this node is an attribute/value pair (selector). The branch from the top root to a leaf is a conjunction of attribute/value pairs (complex). And the leaf is the label of the instances that are covered by this complex. The branch thus is a classification rule.

Growing a decision tree is a recursive process. The pseudo code for growing the tree is shown in Fig. 1. At the beginning, three extreme cases are checked. If any of them happens, it returns a single node tree with a unique leaf. Constant variable

```
R is the set of all attributes
C is the target
S is the dataset
procedure ID3( R, C, S )
    if S=∅ then
        return single node tree with def_label as the leaf;
    end if
    if IsPure(S) then
        return single node tree with Label(S) as the leaf;
    end if
    if R=∅ then
        return single node tree with def_label as the leaf;
    end if
    D ← BestSplit(D,F);
    for all values d_j of all the values of D do
        S_j ← Subset(S, D, d_j);
        return ID3(R-{D}, C, S_j);
    end for
end procedure
```

FIGURE 1. Decision Tree.

def_label means the mostly seen label in the original dataset. And the function label(S) returns the mostly seen label in S. If none of the three happens, the following step is to determine the root attribute. Root attribute is the one that can most accurately classify remaining instances. The attribute with the highest Information Gain is chosen as the root. And the heuristic of information gain is given by Equation (9),

$$G(D, S) = H(D) - \sum_{i=1}^{c} P(D_i)H(D_i), \tag{9}$$

where D is a candidate attribute, S is the data, c is the number of values of attribute D, D_i are the instances with the i^{th} attribute value, $P(D_i)$ denotes the portion of D_i, and $H(D_i)$ is the entropy of D_i. The entropy is given by Equation (10),

$$H(p_1, p_2, \ldots, p_j, \ldots, p_n) = \sum_{j=1}^{n} -p_j \times log_2(p_j), \tag{10}$$

where n is the number of target labels, p_j denotes the portion of instances with the j^{th} label.

The procedure of root selecting is denoted by the function BestSplit. This root extends edges, where each of them denotes a value of this root attribute. For each edge, the instances that are covered by the branch from the top root to this edge (the procedure of finding those instances is denoted by function Subset) are forwarded to a new ID3 procedure for either terminating the branch or growing a new tree. Finally, the decision tree will stop growing if no new trees can be derived.

2) PROBLEM TWO

In order to assign customers tariffs their power consumption features need to be known. In addition to power consumption features, energy suppliers also assign tariffs in another way. They categorise customers into groups according to the shapes of their time-series power consumption curves (i.e., customers with similar consumption patterns are in a same group, and customers with different patterns are in different groups), and then assign different tariffs to different groups of customers. The information of both the power consumption features of a customer or the group that this customer belongs to is derived from customers' power consumption records. For those customers without the records, such information may be estimated based on their dwelling and occupant characteristics.

As discussed by the second question in Section I, if the mapping from customer characteristics to power consumption features (or the group of belonging) can be represented more intuitively, like a manual, then energy suppliers can directly estimate customers' certain power consumption features and assign them proper tariffs without actual mathematical calculation. Therefore, we apply ID3 for classification. Then, rules such as customers in terms of what kinds of characteristics have a high power consumption level, or customers in terms of what kinds of characteristics belong to group one can be discovered. We carry out seven classification sub-problems in total, and all of these sub-problems take customer dwelling and occupant characteristics as attributes. However, these problems have different targets, where each of the first six sub-problems takes a different power consumption feature as the target, and the final sub-problem takes customer's belonging group as the target.

Before carrying out the seventh sub-problem, we need to know, for this Irish ECBT data, which group each customer belongs to. The following describes our procedure for categorising customers into different groups. Firstly, the original half-hourly power consumption records are multiplied by two, which derives customers' loads. Then for each customer we have their daily load profiles (DLPs) of multiple days. A DLP is a 48 element vector denoting the time series of load of the day. We firstly average the multiple DLPs and get the Average DLP (ADLP) for this customer. The ADLP is then divided by the maximum element of this ADLP which obtains the Representative Load Profile (RLP). Finally, we use the Clustering algorithm SOM to cluster customers according to their RLPs. As a consequence, customers in a same group have similar RLPs, and the RLPs of customers in different groups are obviously different. Customers are categorised into 5 groups, so each customer is associated to a group ID.

In Fig. 2, the first five graphs show RLPs of different groups, and graph six shows the RLPs of all customers.

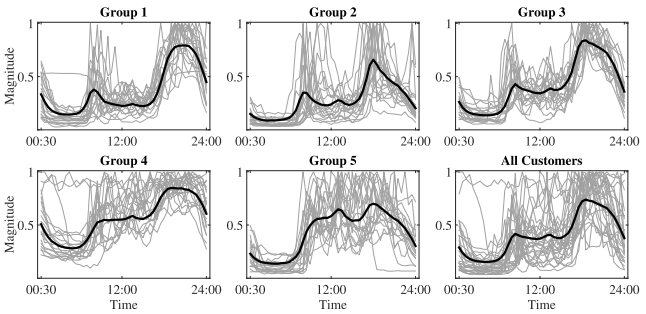


FIGURE 2. Five Categorized Customer Groups Based on RLPs.

In a graph, the averaged RLP of the corresponding group of customers is highlighted with a thick black line, and some examples of individual customers' RLPs are shown as grey lines. It is obvious that power consumption patterns among groups are different. For example, there is a significant difference between the lunch time and the night for group one, but the pattern for group five is more flat. Therefore, energy suppliers can consider assigning group one customers higher night time electricity price and lower lunch time price, which may motivate them to shift the peak power demand. A more scientific method for assigning customers optimal tariffs based on the group averaged RLP is proposed in [26].

We show the detailed data specification for the classification problem in III-D, and the results of all sub-problems in IV-B.

C. SUBGROUP DISCOVERY

Subgroup discovery identifies the complex which covers the statistically most *interesting* subgroup of instances with respect to a chosen target [27]. The term interesting is specified as the trade-off between generality and uniqueness [4]. Generality means the subgroup size, and uniqueness means the difference of the target labels' distribution between the subgroup and the population. For example, provided the proportions of different target labels in a dataset is [40%, 35%, 25%], we may consider a complex that covers 40% instances with a [20%, 35%, 45%] target label distribution to be interesting. A rule of subgroup discovery is in the form of: *Target Distribution* ← *Condition*, where condition is the complex, and target distribution is the distribution of the target labels of all the instances covered by this complex. Applying subgroup discovery for smart meter data is first described in [28], which focuses on problems with numerical target variables. In our study, subgroup discovery serves problems with discrete targets.

To discover the best rule, subgroup discovery evaluates the heuristic scores for all possible complexes. The one with the highest score is discovered. In this study, we use the heuristic of Multi-class Weighted Relative Accuracy (MWRAcc) which is proposed in [22] and mathematically formulates the trade-off between generality and uniqueness:

$$\mathrm{MWRAcc}(b) = \frac{1}{n}\sum_{i=1}^{n} |\mathrm{WRAcc}_i(b)|. \qquad (11)$$

where b is a subgroup and n is the number of target labels. WRAcc is the shorthand of Weighted Relative Accuracy that is proposed in [4] for the binary class subgroup discovery problem. $\mathrm{WRAcc}_i(b)$ is given by: $\mathrm{WRAcc}_i(b) = \frac{e}{E}(\frac{e_i}{e} - \frac{E_i}{E})$, where e is the size of b, e_i is the number of i labelled instances in b, and E and E_i are for the population. CN2-MSD is a subgroup discovery algorithm that is proposed in [22].

1) CN2-MSD

CN2-MSD is adapted from CN2 - a typical classification algorithm [29]. Two main changes adapt CN2 to CN2-MSD. Firstly, their chosen heuristics are different. The heuristic of CN2 is similar to ID3 which is used for classification, while CN2-MSD uses MWRAcc. Secondly, when a rule is found by CN2, it removes all of the covered instances and then uses the remaining instances as the basis to discover another rule. Instead of removing the instances, CN2-MSD assigns weights to those covered instances when a rule is found, and those weights will be taken into account when discovering the next rule.

2) PROBLEM THREE

As discussed in the third question in Section I, we intend to discover the rule that explains customers in terms of what kinds of characteristics will benefit most from DSM for the peak demand shifting purpose. For those customers that use most appliances during the peak time but much less during other times, more appliances can potentially be switched on by DSM at off-peak times, and the subsequent peak shifting effect should be significant. In other words, those customers that use much more electricity in peak time than in off-peak time will benefit most from DSM.

As discussed, all customers have been categorised into five groups according to their power consumption patterns, and their consumption patterns have been presented in Fig. 2. We have also shown the average RLP for the whole population in the sixth graph of this figure. As shown in the sixth graph, we see the peak period starts from 16:00, and the time before is the off-peak period. Among all five groups, group one customers' power consumptions between the peak and off-peak times are most different. And the difference becomes lower and lower with the increase of the group ID. Therefore, our aim is actually to discover the complex that covers as many group one customers as possible but as few group five customers as possible.

This problem is solved by subgroup discovery, in which we use customer group ID as the target and customers' dwelling and occupant characteristics as the attributes. The data specification and the result of this problem are presented in III-D and IV-C respectively.

3) PROBLEM FOUR

The aim is to identify attitudes with which customers consume less electricity. Subsequently, energy suppliers can guide customers to change their attitudes for energy conservation purposes.

This problem is also solved by subgroup discovery with a similar procedure to that for problem three. We firstly categorise customers into five clusters (to avoid confusion, we use the term cluster rather than group for this problem), while the categorisation principle is different. For this case, customers are categorised on the basis of their power consumption amounts rather than the shapes of their power consumption time-series. More specifically, the categorisation is based on customers' ADLPs rather than their RLPs. In Fig. 3, we show the average ADLP of the population and each cluster of customers as different types of lines. It can be seen that the average ADLP of different clusters have similar shapes but

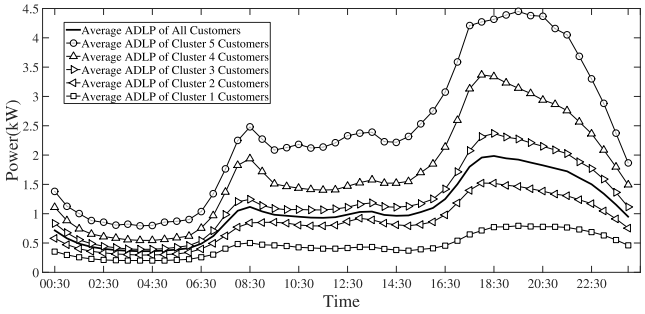


FIGURE 3. Five Categorized Customer Clusters Based on ADLPs.

significantly different magnitudes - the lower the cluster ID is the lower the consumption amount is.

Provided every customer has been associated with a cluster ID, we take the cluster ID as target and customer attitude characteristics referred in II-B-(2) as the attributes to build the data. Finally, subgroup discovery is applied which identifies the complex of attitudes that covers as many low consumption customers (i.e., cluster one customers) as possible but as few high consumption customers (i.e., cluster five customers) as possible. We show the detailed data specification and the result in III-D and IV-D respectively.

TABLE 2. Data specifications for each problem.

Problem ID	Attributes	Discretized Attribute	Target
Problem 1	$\{\mathbf{\Phi}, \mathbf{\Gamma}\}$	Yes	None
Problem 2	$\{\mathbf{\Phi}\}$	No	Γ_i or Group ID
Problem 3	$\{\mathbf{\Phi}\}$	No	Group ID
Problem 4	$\{\mathbf{\Theta}\}$	No	Cluster ID

D. A SUMMARY OF DATA SPECIFICATIONS

In total 30 dwelling and occupant characteristics, 30 attitude characteristics and six power consumption features for 3488 customers, as well as all these customers' Group ID with respect to their load profile shapes and their Cluster ID with respect to their power consumption amounts have been derived. As discussed, the exact data used for the four problems are different, and we show a summary of the data specifications for those problems in Table 2. In the first column, we list all of the four problems. In the second column, we show the data attributes for each problem, where Φ is the set of all dwelling and occupant characteristics, Γ is the set of all power consumption features, and Θ is the set of all attitude characteristics. The third column shows if the numerical attribute variables in a problem needs to be discretized. The final column shows the chosen target for each problem, where Γ_i denotes the i^{th} power consumption feature.

IV. EXPERIMENTS

This section presents the results of the four discussed problems. Apriori, J48 and CN2-MSD are applied, where J48 is a Java implementation of ID3, which handles continuous variables and missing values. We use the implementation of Apriori and J48 provided by the machine learning toolbox WEKA [30]. And for CN2-MSD, we use the implementation provided by [22].

TABLE 3. Association rules discovered by Apriori.

Head	Body	Metric Scores
empl_stat=0	occ_net=0, dwg_own=1	0.18; 0.78; **1.88**; 0.41
empl_stat=0	ent_appl=1	0.19; 0.64; **1.54**; 0.41
income=2	empl_stat=1	0.20; 0.35; **1.47**; 0.23
occ_net=0	ADMD=1	0.17; 0.50; **1.61**; 0.31
occ_net=1	empl_stat=1, education=3	0.25; **0.90**; 1.31; 0.69
occ_net=1	TotalkWh=3	0.15; **0.89**; 1.30; 0.69
occ=1	TotalkWh=1	0.16; 0.39; **1.95**; 0.20
occ=3	ent_appl=3	0.16; 0.48; **1.77**; 0.27
occ=3	empl_stat=1, TOU=1	0.16; 0.43; **1.56;** 0.27
dwg_type=0	bedrooms=3	0.25; **0.73**; **1.34**; 0.54
dwg_type=0	appl=3	0.16; **0.72**; **1.33**; 0.54
dwg_own=0	occ=3	0.21; 0.75; **1.65**; 0.45
dwg_own=0	occ=3, ADLFactor=1	0.11; **0.80**; 1.75; 0.45
dwg_age=1	occ=3	0.21; **0.75**; 1.40; 0.54
dwg_age=1	occ=3, dwg_type=0	0.13; **0.81**; 1.51; 0.54
elec_cook=0	appl=1	0.17; 0.46; **1.53**; 0.30
elec_cook=1	appl=3	0.18; **0.85**; 1.21; 0.70
appl=1	occ=1	0.11; 0.57; **1.51**; 0.38
appl=1	ADMD=1	0.20; 0.59; **1.57**; 0.38
ent_appl=1	occ_net=0	0.20; 0.63; **2.14**; 0.30
ent_appl=1	TotalkWh=1, ADMD=1	0.16; 0.54; **1.83**; 0.30
light_bulbs=1	occ=1	0.14; **0.70**; **1.08**; 0.65
education=3	income=2	0.14; 0.59; **1.63**; 0.36
TotalkWh=1	ADMD=1	0.30; **0.88**; **2.15**; 0.41
TotalkWh=2	ADMD=2	0.30; 0.67; **1.60**; 0.42
ADMD=1	ent_appl=1	0.18; 0.61; **1.77**; 0.34
ADLFactor=1	TOU=1, EveImpact=3	0.11; **0.77**; **1.55**; 0.50
ADLFactor=1	elec_cook=1, EveImpact=3	0.11; **0.79**; 1.59; 0.50
TOU=1	EveImpact=3	0.14; **0.74**; 1.26; 0.59
LunImpact=1	EveImpact=3	0.16; **0.84**; **1.81**; 0.46

A. PROBLEM ONE RESULTS

For this problem, three types of association rules are discovered, which represent the relations among dwelling and occupant characteristics, the relations among power consumption features and the relations between the two kinds of variables. As shown in Table 3, column one and two store these rules. As discussed, confidence and lift are the knowledge searching heuristics and we also use both as the metrics for evaluating the qualities of the discovered rules. A cell of the third column is a four element score vector where the first three elements are the support, the confidence and the lift of a rule, and the fourth element is the support of this rule's head. Let us recall that the confidence score of a rule denotes the possibility of the head given the body, so: the higher the score is, the safer it is to imply the head provided the body occurred. The lift score is the ratio of the observed support of a rule to the expected support when this rule's body and head are independent, so: the higher the score is, the higher the dependency is. The support of a rule's head is also provided for justifying the dependency. All of the rules in this table are discovered by either confidence or lift. To distinguish them, a rule's confidence score is printed in bold if it is found by confidence. Rules found by lift are highlighted in the same way. In the following, we show some examples for each type of association rules.

We start by discussing those rules that are about dwelling and occupant characteristics. Rule {occ_net=1 ← empl_stat=1, education=3} means the probability that there is regular Internet user in a dwelling is 90% if the chief income earner of the dwelling is an employee with high educational level. The confidence score of 90% is much higher than the head support of 69% which implies a high dependency which is also proved by the 1.31 lift score. Another rule {dwg_own=0 ← occ=3} is found by lift which indicates a high dependency between the case that both adults and children live in the dwelling and the other case that the dwelling is either rented or owned but with mortgage. We find the average daily consumption of the customers owning outright (dwg_own=1) is 23.6 kWh, and the average consumption of the other customers which either rent their home or own with mortgage (dwg_own=0) is 26.8 kWh. The higher consumption of the latter customers is probably because their dwelling usually has a large number of people (occ=3).

We now show associations of power consumption features. Rule {TotalkWh=1 ← ADMD=1} and {TotalkWh=2 ← ADMD=2} show the total power consumption is highly correlated with the maximum power demand, which claims the importance of appliance rating for energy conservation. Another rule {LunchImpact=1 ← EveImpact=3} indicates a 84% conditional probability that: the lunch time power consumption is low given that the evening time power consumption is high. The lift score is as high as 1.81 which claims the head is highly dependent on the body. Low lunch time consumption can be caused by the absence of occupants, so they seem to have day time jobs. Such occupants are usually at home during evening time which raises the power consumption.

Associations between power consumption features and customer characteristics are more interesting. Rule {occ_net=1 ← TotalkWh=3} indicates the probability that there is regular Internet user is 89% given that the total power consumption of this dwelling is high. Therefore, it will be effective to get in touch with the high total power consumption customers via Internet like email (e.g., for effectively sending feedback to customers). Another rule {occ_net=0 ← ADMD=1} denotes an opposite situation that if the averaged daily maximum demand of a household is low then there is a 50% probability that all the occupants do not use Internet regularly. For this case, these customers are better to be contacted by post rather than email.

In summary, more associations like the one that is referred in the first question in Section I can be discovered by association rule learning algorithms. We have shown that the discovered rules are not only mathematically interesting (i.e., high confidence and lift) but also useful.

B. PROBLEM TWO RESULTS

We have conducted seven classification sub-problems which are distinguished by the chosen target. The target of a sub-problem is either one of the six power consumption features or the group ID of Fig. 2. In this section we present the representative rules and the accuracy for each sub-problem.

The ultimate output of a classification sub-problem is a tree. We take the tree of the TotalkWh sub-problem as an example to describe how energy suppliers classify a customer with respect to a target. The tree is shown in Fig. 4. A circle is an attribute which can either be *occ* (the number of occupants) or *appl_use* (How often are the appliances used). An edge extended from a circle specifies a bound of the corresponding attribute. A leaf (square) is a TotalkWh label, which is 1 (low), 2 (medium) or 3 (high). The joint attribute/value pairs and the label along the branch from the top circle to a leaf induces a rule. The rule covered instances' number and the number of instances that are misclassified are seen in the leaf of this branch as well.

The left most branch induces rule {TotalkWh=1 ← occ≤1} which implies that the total power consumption of a customer is low if there is only one person living in the house. The accuracy of the rule is calculated by: $\frac{743-163}{743}$ which equals 78%. An opposite rule is from the right most branch, which leads to the customers of high total power consumption. It is {TotalkWh=3 ← occ>3, appl_use>12} which implies that if the dwelling has more than 3 people and the appliances are used more than 12 times per day, the total power consumption of this dwelling is likely to be high. This rule is with a 60% classification accuracy. Both rules seem rational, because the more occupants there are, the more frequently appliances are used and the higher the total power consumption is. For the other classification sub-problems, we are not presenting their decision trees,

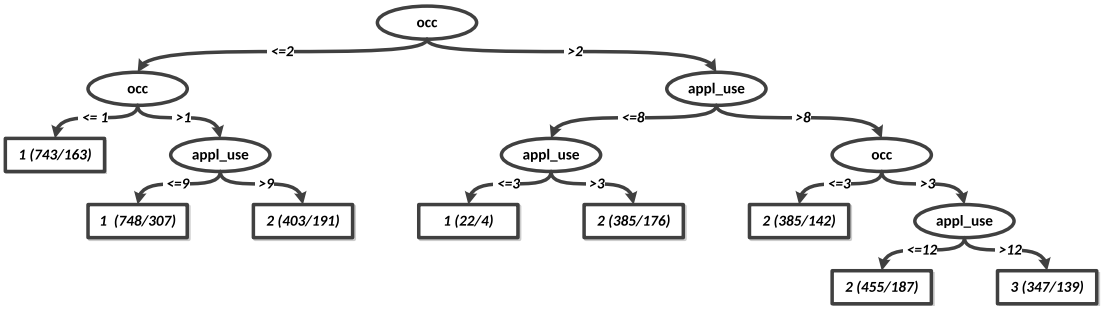


FIGURE 4. The Decision Tree with respect to TotalkWh.

but only show some example rules. Each example rule is followed by brief discussions as to why this rule is reasonable. The accuracies of the following rules are 78.6%, 73.4%, 63.2%, 57.0%, 54.8% and 55% respectively.

Rule-1: If the number of occupants is less than 2 and the appliances' times of use is less than 8, then the averaged daily maximum demand is low. As we described in IV-A, averaged daily maximum demand is highly correlated with total power consumption; it is thus proportional to the occupants' counts and appliances' times of use as well.

Rule-2: If the interviewee is younger than 56, the number of bedrooms is less than 4, electricity is used for cooking, and entertainment appliances' times of use is less than 13, then the averaged daily load factor is low which means the power consumptions are not stable. Using electricity for cooking increases the power consumption. Besides, people younger than 56 are more likely employed, so they are not home until dinner time. Therefore, the power consumption of a day is not stable.

Rule-3: If the number of entertainment appliances is more than 2, the customer is a peak-time user. This rule is not significant. In the following, we will see TOU cannot be classified accurately.

Rule-4: If the chief income earner is unemployed, then the lunch impact is medium. Such occupants are likely staying at home during the daytime, thereby their lunch time power consumptions are relatively higher than the other customers.

Rule-5: If the interviewee is older than 35 but younger than 55, and there is at least one adult person staying at home during the daytime, then the evening impact is low. It is possible that the occupant that stays at home during the daytime can prepare dinner or have other activities that consume electricity before evening, so the evening consumptions are relatively low.

Rule-6: If the interviewee is younger than 36, and there is no one staying at home during the daytime, then the customer belongs to group one. The occupants are likely employed, and there is no one staying at home during the daytime, so we can see in the first graph of Fig. 2 that the averaged RLP is low during the daytime but high in the evening.

Furthermore, we assess these seven classification sub-problems' performances in terms of their classification accuracy. We show statistics of accuracy in Fig. 5 with the x-axis listing all the seven sub-problems and the y-axis denoting accuracy. For each sub-problem, we show the classification accuracy by J48, and also the accuracies by two other reference classifiers which are random guess and biased random guess respectively. Using the two reference classifiers for evaluating the result of J48 is proposed in [31]. In a sub-problem with K target labels, the accuracy of random guess is $\frac{1}{K}$, because customers are assigned with labels uniformly. Since biased random guess assumes the distribution of the target labels is known, its accuracy is thus given by $\sum_{k=1}^{K}(\frac{S_k}{S})^2$, where S is the size of population, and S_k is the number of instances labelled with k. The provided accuracies are all from 10-fold validation.

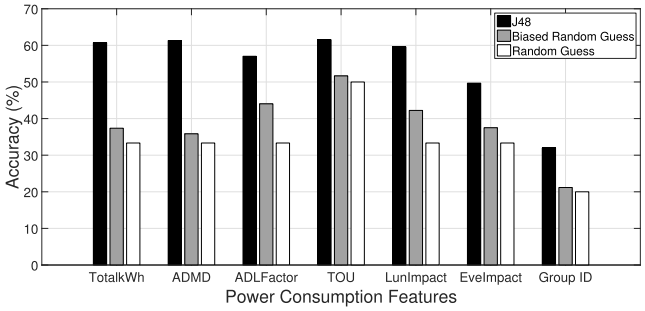


FIGURE 5. Classification Accuracies of the 7 Sub-problems.

J48 accuracies for TotalkWh, ADMD and LunchImpact approach 60%, which are significantly higher than their random guess accuracies 38%, 37% and 42% respectively. Therefore, given customers' characteristics, energy suppliers can estimate these three power consumption features with good accuracy. The J48 accuracy on ADLFactor is not as good, but is still higher than the reference classifiers'. Besides, for ADLFactor, the accuracies for different target labels are found not even: customers labelled with 1 are classified more accurately (i.e., only $\frac{1}{3}$ label 1 customers are misclassified). In other words, customers with low averaged daily load factor can still be recognised based on their characteristics. The rest of variables cannot be accurately classified.

In summary, our evaluation shows certain power consumption features like the total power consumption, averaged daily maximum demand and lunch time impact can be classified accurately based on customer dwelling and occupant characteristics. Energy suppliers may thus consider using these features as a basis when assigning customers tariffs. Besides, ID3 is able to represent the classification paradigms as a tree-like graph, which is intuitive and easily understandable. Furthermore, showing energy suppliers classification rules can facilitate them understanding how a target label is caused by a certain number of customer characteristics.

C. PROBLEM THREE RESULTS

We discovered the rule that explains customers in terms of what kinds of dwelling and occupant characteristics can benefit most from DSM. The body part of this rule is described as: {occ+ $\leq$ 3, occ- $\leq$ 3, voice_age $\leq$ 4, home_occ $\leq$ 1, empl_stat = 1, ent_use $\leq$ 8, ent_appl $\leq$ 7, appl_use $\leq$ 15, dwg_age $\geq$ 2, bedrooms $\geq$ 2, appl $\leq$ 10}, which covers 16% of the total customers.

On the left of Fig. 6, we show the comparison of group ID distribution between the subgroup and the population. It can be seen that the group one customers' proportion in the subgroup is significantly increased, but the fifth customer group has an opposite situation. To highlight the uniqueness of the subgroup relative to the population, we also show the Relative proportion difference (RPD) for each group on the right of this figure, which is given by: $RPD_i = \frac{p_i - Pi}{Pi}$,

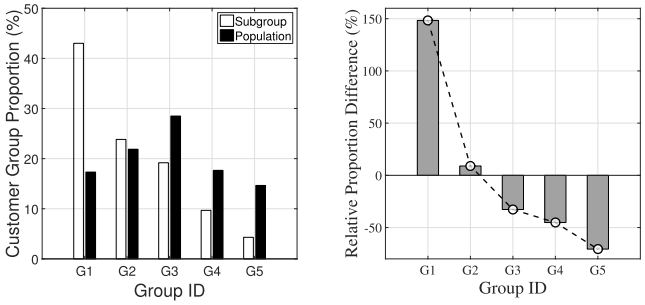


FIGURE 6. Problem Three Subgroup. The graph on the left shows the Group ID distributions of the subgroup and the population, and the graph on the right are the relative proportion differences between the subgroup and the population.

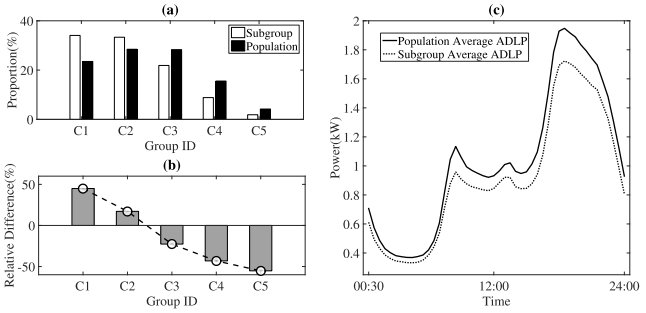


FIGURE 7. Problem Four Subgroup. a, cluster ID distributions of the subgroup and the population; b, the relative proportion differences of the subgroup to the population; c, the effect of customer attitudes on energy conservation.

where i is the group ID, and p_i and P_i denote the proportion in the subgroup and in the population respectively. A larger RPD_i indicates that the complex has strong intention to change the proportion of the i^{th} group of customers. It can be seen that the RPD of group one nearly approaches 150%, and it continuously decreases with increased group ID. In other words, the complex intends to cover as many group one customers as possible but remove as many group five customers as possible. Therefore, we claim this complex intends to identify customers that will benefit most from DSM.

The complex explains why the uniqueness of this subgroup happens. Two attribute/value pairs draw our interest which are *home_occ≤1* and *empl_stat=1*. They mean there is usually no more than one person staying in the dwelling during the daytime, and the chief income earner of the dwelling is employed. Both characteristics imply low power consumptions during daytime. Therefore, installing DSM will make significant improvement for those dwellings where at most one person is usually staying during daytime and the chief income earner is employed.

D. PROBLEM FOUR RESULTS

The data is split into two parts of 2000 and 1488 customers respectively. We learn the rule about customers in terms of what kinds of attitudes consume less electricity based on the first part of the data. We then use the second part to evaluate if the changes of customers' attitudes in the way our rule guided can save energy.

The body part of our discovered rule includes the following four attitudes:

- Att_1*:* if the customer thinks it is too inconvenient to reduce the electricity usage;
- Att_2*:* if the customer thinks he/she is not able to get the people he/she lives with to reduce their electricity usage;
- Att_3*:* what the customer thinks of the opportunity to sell back extra electricity to his/her electricity supplier;
- Att_4*:* how the customer thinks that his/her electricity bills will change as part of trial.

Both values of Att_1 and Att_2 are on a scale of 1 to 5 where 1 is strongly agree and 5 is strongly disagree. The value of Att_3 is on the same scale where 1 is very satisfied and 5 is very dissatisfied. The value of Att_4 is on a scale of 0 to 6 which denote increase, no change, decrease less than 5%, decrease 5%-10%, decrease 10%-20%, decrease 20%-30% and decrease more than 30% respectively. And the exact complex that we found is $\{Att_1 \geq 2, Att_2 \geq 3, Att_3 \leq 4, Att_4 \leq 3\}$.

We show the comparison of cluster ID distribution between the subgroup and the population on graph (a) of Fig. 7. The proportions of both cluster one and two customers in the subgroup are increased with respect to the population, but the proportions for the rest of the clusters are decreased. We also show the Relative proportion difference in graph (b). It is clear that with the increase of the cluster ID, the customer proportion continuously decreases. Let us recall that the higher the cluster ID is, the higher the power consumption is, so this rule intends to remove more high consumption customers but retain more low consumption customers.

Let us turn back to explain why low consumption happens to this subgroup of customers. $Att_1 \geq 2$ means these customers do not have the extreme thought that saving electricity usage is too inconvenient. $Att_2 \geq 3$ means these customers relatively agree that they can get other people in the house to reduce their electricity usage. Both attitudes reflect these customers have a certain belief in their capability of saving energy. $Att_3 \leq 4$ means these customers do not care about the chance to sell back extra electricity at all, which implies these customers are not interested in reducing their electricity bill. $Att_4 \leq 3$ denotes these customers do not have ambitious goals of electricity bill saving, probably because these low consumption customers had the experience of saving electricity before and found it is not that easy.

As a consequence, energy suppliers can try to guide customers to change their attitudes via feedback provision, in order to encourage them to save energy. For example, energy tips should be provided to let customers know some energy saving activities are not inconvenient. Besides, energy suppliers can mildly encourage customers and convince them that they can get other people in their household to reduce electricity. Furthermore, energy suppliers can manage customers' expectations as to how much they will save on

their electricity bills, to avoid customers developing overly ambitious expectations by themselves.

Based on the second part of the data, we found the subgroup of customers that satisfy the above rule. We show the average ADLP of the subgroup and the population in Fig. 7. Their difference reflects to what extent can customers' electricity be saved if customers' attitudes are altered in the way our rules guided. The average ADLP of the subgroup of customers is significantly lower than the population's. The average daily power consumption of the subgroup of customers is 22.03 kWh, and the population's is 24.91 kWh. Therefore, customers with the attitudes that we described above can save 11.6% electricity on average.

V. CONCLUSION

Rule induction is a practical approach of knowledge discovery. Provided a problem is developed, the rule induction technique is able to discover knowledge that addresses the goal of this problem automatically, and the technique is featured by the way of the knowledge representation. In this paper, four energy efficiency related problems are solved by three rule induction techniques, namely classification rule learning, association rule learning and subgroup discovery. We have shown the potential of rule induction for energy efficiency.

We have used the Apriori algorithm to discover knowledge that is mathematically interesting. Our results show some of the discovered rules are not only interesting but also useful. Some rules facilitate energy suppliers to understand customers' certain consumption features. For example, one rule implies if both adults and children live in a dwelling, then this dwelling is likely a rented one or owned but with mortgage. In other words, there are usually more people in such dwellings. This explains why customers in rented dwellings or with mortgage of their dwellings usually use more electricity than those customers that own their dwellings outright. Some rules enable energy suppliers to improve services, e.g., one rule implies high power consumption customers are more likely regular Internet users, so energy suppliers can send such customers feedback by email.

We have used J48 to classify customers' power consumption features based on their dwelling and occupant characteristics. Our results show total power consumption, lunch time impact and averaged daily maximum demand can be estimated accurately, so the three can be used when assigning customers tariffs. Furthermore, J48 enables the classification model to be interpreted as a set of predictive rules, based on which a customer's consumption features can be estimated without any calculation. Besides, each rule gives insight as to how a consumption feature is caused by certain characteristics.

By using CN2-MSD, we have identified those households that would benefit most from DSM. For these households, the chief income earner is an employee, and at most one person usually stays at home during the daytime. Energy suppliers can thus consider installing DSM at these households first. CN2-MSD has also identified the customer attitudes that dominate energy conservation. The results show enhancing customers' belief on their energy saving capabilities and informing customers how much energy they can save can reduce their energy consumption. If all customers' attitudes are changed in the way that our rules guide, nearly 11.6% energy can be saved.

ACKNOWLEDGMENT

The authors wish to take this opportunity to thank the Commission for Energy Regulation (CER) for supplying the Electricity Customer Behavior Trial database.

REFERENCES

[1] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Mag.*, vol. 17, no. 3, pp. 37–54, 1996.

[2] S. Greco, B. Matarazzo, R. Slowinski, and J. Stefanowski, "An algorithm for induction of decision rules consistent with the dominance principle," in *Rough Sets and Current Trends in Computing* (Lecture Notes in Computer Science), W. Ziarko and Y. Yao, Eds. Berlin, Germany: Springer-Verlag, 2001, pp. 304–313.

[3] J. Fürnkranz, "Separate-and-conquer rule learning," *Artif. Intell. Rev.*, vol. 13, no. 1, pp. 3–54, 1999.

[4] N. Lavrač, B. Kavšek, P. Flach, and L. Todorovski, "Subgroup discovery with CN2-SD," *J. Mach. Learn. Res.*, vol. 5, pp. 153–188, Feb. 2004.

[5] I. MacLeay *et al.*, "Digest of United Kingdom energy statistics 2014," Dept. Energy Climate Change, London, U.K., Tech. Rep. ISBN 9780115155307, 2014.

[6] R. de Sá Ferreira, L. A. Barroso, P. R. Lino, M. M. Carvalho, and P. Valenzuela, "Time-of-use tariff design under uncertainty in price-elasticities of electricity demand: A stochastic optimization approach," *IEEE Trans. Smart Grid*, vol. 4, no. 4, pp. 2285–2295, Dec. 2013.

[7] S. Gottwalt, W. Ketter, C. Block, J. Collins, and C. Weinhardt, "Demand side management—A simulation of household behavior under variable prices," *Energy Policy*, vol. 39, no. 12, pp. 8163–8174, 2011.

[8] G. Chicco, R. Napoli, P. Postolache, M. Scutariu, and C. Toader, "Customer characterization options for improving the tariff offer," *IEEE Trans. Power Syst.*, vol. 18, no. 1, pp. 381–387, Feb. 2003.

[9] F. McLoughlin, A. Duffy, and M. Conlon, "Characterising domestic electricity consumption patterns by dwelling and occupant socio-economic variables: An Irish case study," *Energy Buildings*, vol. 48, pp. 240–248, May 2012.

[10] T. A. Nguyen and M. Aiello, "Energy intelligent buildings based on user activity: A survey," *Energy Buildings*, vol. 56, pp. 244–257, Jan. 2013.

[11] S. Darby, "The effectiveness of feedback on energy consumption. A review for Defra of the literature on metering, billing and direct displays," Environ. Change Inst., Oxford, U.K., Apr. 2006.

[12] H. Allcott, "Social norms and energy conservation," *J. Public Econ.*, vol. 95, nos. 9–10, pp. 1082–1095, 2011.

[13] H. S. Cho, T. Yamazaki, and M. Hahn, "AERO: Extraction of user's activities from electric power consumption data," *IEEE Trans. Consum. Electron.*, vol. 56, no. 3, pp. 2011–2018, Aug. 2010.

[14] Y. G. Yohanis, "Domestic energy use and householders' energy behaviour," *Energy Policy*, vol. 41, pp. 654–665, Feb. 2012.

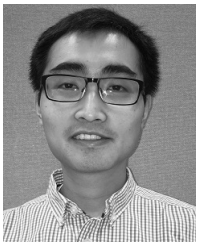
[15] A. Nilsson, C. J. Bergstad, L. Thuvander, D. Andersson, K. Andersson, and P. Meiling, "Effects of continuous feedback on households' electricity consumption: Potentials and barriers," *Appl. Energy*, vol. 122, pp. 17–23, Jun. 2014.

[16] I. Vassileva and J. Campillo, "Increasing energy efficiency in low-income households through targeting awareness and behavioral change," *Renew. Energy*, vol. 67, pp. 59–63, Jul. 2014.

[17] W. Abrahamse and L. Steg, "How do socio-demographic and psychological factors relate to households' direct and indirect energy use and savings?" *J. Econ. Psychol.*, vol. 30, no. 5, pp. 711–720, 2009.

[18] C. Beckel, L. Sadamori, T. Staake, and S. Santini, "Revealing household characteristics from smart meter data," *Energy*, vol. 78, pp. 397–410, Dec. 2014.

[19] Commission for Energy Regulation (CER). *Smart Meter Electricity Trial Data*. [Online]. Available: http://www.ucd.ie/issda/data/commissionforenergyregulationcer/, accessed May 10, 2015.
[20] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Proc. 20th Int. Conf. Very Large Data Bases (VLDB)*, 1994, pp. 487–499.
[21] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.
[22] T. Abudawood and P. Flach, "Evaluation measures for multi-class subgroup discovery," in *Machine Learning and Knowledge Discovery in Databases* (Lecture Notes in Computer Science), vol. 5781, W. Buntine, M. Grobelnik, D. Mladenić, and J. Shawe-Taylor, Eds. Berlin, Germany: Springer-Verlag, 2009, pp. 35–50.
[23] V. Figueiredo, F. Rodrigues, Z. Vale, and J. B. Gouveia, "An electric energy consumer characterization framework based on data mining techniques," *IEEE Trans. Power Syst.*, vol. 20, no. 2, pp. 596–602, May 2005.
[24] S. V. Verdú, M. O. García, C. Senabre, A. G. Marín, and F. J. G. Franco, "Classification, filtering, and identification of electrical customer load patterns through the use of self-organizing maps," *IEEE Trans. Power Syst.*, vol. 21, no. 4, pp. 1672–1682, Nov. 2006.
[25] M. Vannucci and V. Colla, "Meaningful discretization of continuous features for association rules mining by means of a SOM," in *Proc. ESANN*, Jun. 2006, pp. 489–494.
[26] N. Mahmoudi-Kohan, M. P. Moghaddam, and M. Sheikh-El-Eslami, "An annual framework for clustering-based pricing for an electricity retailer," *Electr. Power Syst. Res.*, vol. 80, no. 9, pp. 1042–1048, 2010.
[27] W. Klösgen, "Explora: A multipattern and multistrategy discovery assistant," in *Advances in Knowledge Discovery and Data Mining*, U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds. Menlo Park, CA, USA: American Association for Artificial Intelligence, 1996, pp. 249–271.
[28] N. Jin, P. Flach, T. Wilcox, R. Sellman, J. Thumim, and A. Knobbe, "Subgroup discovery in smart electricity meter data," *IEEE Trans. Ind. Informat.*, vol. 10, no. 2, pp. 1327–1336, May 2014.
[29] P. Clark and T. Niblett, "The CN2 induction algorithm," *Mach. Learn.*, vol. 3, no. 4, pp. 261–283, 1989.
[30] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *ACM SIGKDD Explorations Newslett.*, vol. 11, no. 1, pp. 10–18, Jun. 2009.
[31] C. Beckel, L. Sadamori, and S. Santini, "Automatic socio-economic classification of households using electricity consumption data," in *Proc. 4th Int. Conf. Future Energy Syst. (e-Energy)*, 2013, pp. 75–86.



QIPENG CHEN received the B.Sc. degree in computer science from the University of Ulster, Ulster, U.K., in 2010, and the M.Sc. degree in advanced computing from the University of Bristol, Bristol, U.K., in 2011, where he is currently pursuing the Ph.D. degree with the Electrical and Electronic Engineering Department. His research interests include smart grid, machine learning, and data mining.

ZHONG FAN received the B.S. and M.S. degrees in electronic engineering from Tsinghua University, China, and the Ph.D. degree in telecommunication networks from Durham University, U.K. He was a Research Fellow with Cambridge University, a Lecturer with Birmingham University, and a Researcher with Marconi Laboratories, Cambridge. He is currently a Chief Research Fellow with Toshiba Research Europe, Bristol, U.K. His research interests are wireless networks, IP networks, M2M, and smart grid communications. He received a BT Short-Term Fellowship for his work at BT Laboratories.

DRITAN KALESHI received the Ph.D. degree from the University of Bristol, and the Dipl.Ing. (Hons.) degree in electronics from the Polytechnic University of Tirana, Albania. He was a Senior Lecturer of Communication Networks with the University of Bristol until 2015. He is currently a Visiting Research Fellow with the University of Bristol, and a 5G Fellow with Digital Catapult, London, U.K. He has authored over 60 papers, edited two international standards, and holds three patents. His research interests are in future networking architectures and protocols (5G and beyond), large scale loosely-coupled distributed systems design, modelling and performance evaluation, and data interoperability for sensor/actuator systems (IoT). He represents the U.K. in various international standardisation bodies (ISO/IEC, CEN/CENELEC) in areas related to IoT, home electronic systems, and smart grid.

SIMON ARMOUR received the B.Eng. degree from the University of Bath, Bath, U.K., in 1996, and the Ph.D. degree from the University of Bristol, Bristol, U.K., in 2001.

He has been a member of the Academic Staff with the University of Bristol, since 2001, where he has been a Senior Lecturer since 2007. He has authored or co-authored over 100 papers in the field of baseband PHY layer and MAC layer techniques for wireless communications with a particular focus on orthogonal frequency-division multiplexing, coding, multiple input/multiple output, and cross-layer multiuser radio resource management strategies. He has investigated such techniques in general terms, and specific applications to 802.11, 802.16, and 3GPP LTE.