Socio-Demographic Characteristics Prediction using Soft Clustering of Load Consumption Data

Chung Ming Cheung, Sanmukh Rao Kuppannagari, and Viktor K. Prasanna

University of Southern California, 3740 McClintock Ave, CA90089, USA,
chungmin@usc.edu, kupanna@usc.edu, prasanna@usc.edu

Abstract. Understanding socio-demographic characteristics of customers is important for utility companies to design better policies for demand response programmes. Manual collection of such characteristics via methods like surveys are costly and is prone to errors or unavailability of complete data. Unknown or missing characteristics can be predicted using features extracted from electricity consumption data. Typically, these extracted features are statistical figures related to the electricity usage in specific time periods. In this paper, we propose the use of membership vectors from soft clustering of electricity consumption time series as features for prediction of socio-demographic characteristics. The membership vector indicates how similar each customer is to every consumption pattern cluster. By discovering the correlation between clusters and characteristics, more accurate prediction of characteristics can be performed. Our experiments on a real survey dataset show that the combination of using statistical features of consumption and clustering membership vector as input features gives better classification accuracy than solely using either type of features.

Keywords: soft clustering, time series, machine learning, socio-demographic features

1 Introduction

Socio-demographic characteristics of customers include personal information such as age and ethnicity. They also include socio-economic variables like the employment status and annual income of the customers. In the case of electricity distribution grids, characteristics regarding the household and electricity usage are also considered. This includes characteristics like age and size of the house, resident occupancy, and usage patterns of high consumption electric appliances. It is desirable for utility companies to acquire this data as with the knowledge of these characteristics, the utility can have a better understanding of the customer behavior and demand profiles [1]. This is especially useful for demand response programs [2] as utilities can provide the right incentive for customers to adjust their consumption [3].

To acquire socio-demographic characteristics, utilities carry out customer surveys. However, this method is not reliable for acquiring the characteristics of all customers. Firstly, this is a costly method in terms of both time and human resource as it requires manually conducting surveys through means like telephone calls. Moreover, many customers refuse to respond to questions in the survey due to personal reasons like privacy. This is especially true for sensitive information like household income. For example, in the CER dataset [4], we can see that only 45.6% of annual income data is available, while some customers refused to answer, others only provided a rough weekly income estimation.

To discover the missing socio-demographic characteristics, the correlation between these characteristics and the electricity consumption of the customers can be utilized. The proliferation of smart metering instruments in electricity distribution grids have enabled extensive measurement of customer loads. This allows access to historical electricity consumption time series of each customer at fine granularity time intervals [5]. By discovering the correlations between the electricity usage patterns and customer characteristics, we can then use known electricity usage patterns to predict unknown customer characteristics.

Several approaches can be used to perform customer characteristics prediction. One approach is to perform user profiling to identify key consumption profiles, and analyze the correlation between each type of profile to customer characteristics [6]. It is also possible to use data-driven models to learn a function to map electricity consumption characteristics to customer features outputs. For this approach, it is important to discover relevant features to extract from consumption time series to be used as inputs of the data-driven model [7]. This is because time series data are very high dimensional and not amenable to be used directly inputs due to the curse of dimensionality [8]. Existing methods of feature extraction include manually defining statistical features related to consumption figures or automatically learning low dimension features from time series through data-driven architectures like autoencoders. The former method requires knowledge of meaningful statistical figures to extract, while the latter method is a blackbox method that extracts features that is not comprehensible by humans.

In this paper, we propose to combine clustering and data-driven approaches by defining a new kind of input features based on clustering results. While normal clustering methods assign each data point to one cluster, soft clustering methods produce a membership vector for each data point that describes its similarity to each cluster. This membership vector can be used as one of the input features of data-driven models for the prediction of household socio-demographic characteristics. We explore the performance of using clustering features for characteristics prediction using a real life dataset. We consider two sets of features from consumption data, a set of statistical features computed from the consumption values in different periods of time of a day, and a set of clustering features based on soft clustering. Then, we compare the accuracy of prediction models trained on either set of features or both set of features combined.

Our contributions are summarized as follows:

- We propose the use of membership vectors obtained from soft clustering for the prediction of socio-demographic characteristics of households.

- We show that the performance of socio-demographic characteristics prediction is improved by using a combination of both statistical features and clustering features compared to solely only using one set of features.

The rest of the paper is organized as follows. In Section 2, we describe existing research that is related to household characteristics prediction. In Section 3, we give a formal problem definition of the household characteristic prediction problem that we solve in this paper. In Section 4, we describe our proposed methodology in detail. In Section 5, we explain the experimental setup and analyze the results of the experiments. In Section 6, we propose future directions that can be pursued. Finally, we conclude the paper in Section 7.

2 Related Work

Existing methods for socio-demographic characteristics prediction can be divided into two categories: manual feature extraction approaches and clustering approaches.

Manual feature extraction methods predict the household characteristics using statistical features regarding the consumption figures extracted from load [7, 9]. These features include the maximum or minimum load within particular time periods of the day, ratios between certain statistics, etc. While these methods are able to predict certain characteristics with high accuracy, they do not work well for all characteristics.

Clustering approaches refer to the use of fuzzy clustering to group load consumption time series that exhibit similar behavior. Fuzzy clustering refers to clustering approaches that assign each data point to one or more clusters. The clustering can then be used to discover correlations between clusters of customers. They can then form fuzzy-based rules to predict characteristics for members of each cluster. Examples of works that take the clustering approach are [6, 10].

The advantages of clustering approaches are that they are unsupervised, as clustering can be done using any black box clustering methods. The fuzzy-based rules formed can also provide insights on how different characteristics may correlate, for example, certain demographics may show similar electrical appliance usage patterns. On the other hand, manual feature extraction methods require human expertise in designing features for the prediction model. The advantage of such models over clustering approaches is that it is able to consider more complex temporal correlations in the time series. For example, considering the ratio of electricity usage between certain time periods.

In this work, we investigate the use of fuzzy clustering results as a kind of input feature in addition to other features extracted directly from load to enhance characteristics prediction model accuracy. We believe that this approach can combine the advantages of both methods to produce more accuract characteristics prediction.

3 Problem Definition

In this problem, we assume that we have the electricity consumption historical time series data for all customers, while socio-demographic characteristics data is available for a subset of all customers. Electricity consumption time series is a sequential data of the power consumption of each customer over time. Socio-demographic characteristics are features related to the house, the location and its inhabitants, a more detailed discussion is provided in Section 5.

Using the available socio-demographic features, a prediction model is trained to predict each feature from the consumption data. In this paper, we consider the problem where all socio-demographic features are represented as a binary value, that is, each value represents that a certain property is either "True" or "False" (e.g. The income of the household is above a certain threshold).

Given N customers electricity consumption time series $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ and their socio-demographic data $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N\}$, where $\mathbf{y}_i = \{y_i^{(1)}, y_i^{(2)}, ..., y_i^{(D)}\} \in R^D$ is a D-length vector of binary values representing socio-demographic features of the i-th customer. We extract input features from electricity consumption $\mathbf{X}$ for each customer to get $\mathbf{h} = \{\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_N\}$. The goal is to train a predictor function f_d for each socio-demographic feature d such that it predicts the binary value of the d-th feature for the i-th customer $\hat{y}_i^{(d)} = f_d(\mathbf{h}_i)$. The predictor should minimize the error between the predicted binary values $\hat{\mathbf{y}}_i$ and the actual socio-demographic data $\mathbf{y}_i$.

We make the following assumptions in this problem:

- All socio-demographic characteristic features are binary values. This assumption simplifies the predictor model requirements, allowing it to be trained more quickly and efficiently. If a more detailed characteristic is required (e.g. we need to know the exact number of residents instead of if the number of residents is above a certain threshold), the concerned feature can be extended into several binary problems (e.g. for the number of residents feature, have several binary problems with different thresholds). Thus, considering only binary values as prediction outputs does not limit the capability of the method.
- The electricity consumption data is fully available and there is no missing values in the data. Missing values in data can be recovered via missing data imputation methods [11]. We assume all consumption data is available for simplicity.

4 Methodology

The overall workflow for household characteristics prediction can be summarized as follows. Firstly, statistical features are extracted from electricity consumption time series. Daily load profiles are prepared and preprocessed from the consumption data, which are then used for performing soft clustering to obtain clustering features. A prediction model is then trained using the prepared features to predict each household characteristic. Figure 1 illustrates the workflow for feature

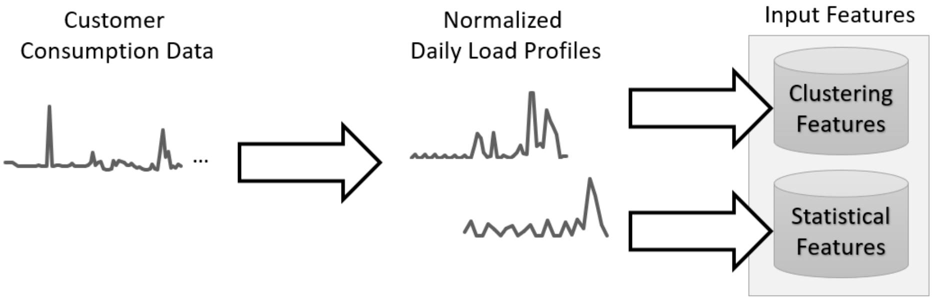

Fig. 1. Workflow for feature extraction from customer consumption data

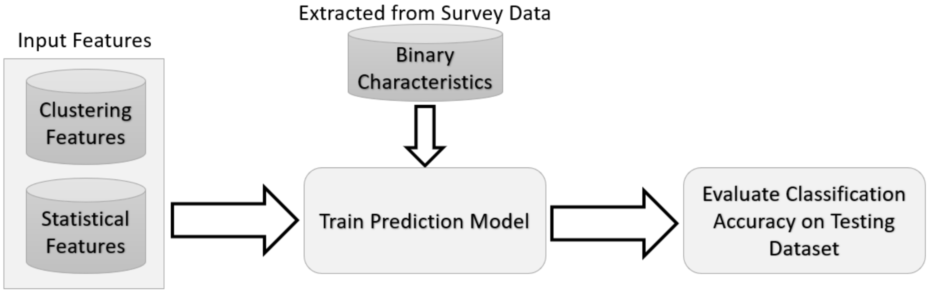

Fig. 2. Workflow for training and evaluating the characteristics prediction model

extraction from customer consumption data, and Figure 2 summarizes the training and evaluation procedure of the characteristics prediction model.

Soft Clustering We use SImilarity-Based Soft Clustering (SISC) [12] for soft clustering. SISC is a generic centroid-based soft clustering algorithm for clustering data points, meaning that each data point can belong to more than one cluster. It is centroid-based because it uses the distance between centroids and each data point as guidance to updating the clusters. The algorithm starts by initializing each cluster by assigning one data point as the centroid of each cluster. Then it iteratively recalculates the similarity between each data point and centroid to determine if a data point belongs to a cluster. This is repeated until a maximum number of iterations is reached or there is no change in assignments. To apply SISC for time series soft clustering, we need to use a suitable similarity measure for time series.

To compute the distance between time series, we use Normalized Cross-Correlation (NCC) similarity proposed by [13]. Traditional distance measures like Euclidean distance is not suitable for measuring time series distances, because such measures are not shift-invariant. Shift-invariance is needed because the distance between any two time series that is identical but shifted by a short time interval would have a great distance. This is undesirable in time series analytics as even if time series are slightly shifted, we are more concerned with their shape in terms of similarity. On the other hand, dynamic time warping (DTW) [14] solves this shift-invariance problem by allowing each data point on a time series to be compared to any data point within a certain time window on the other time series. However, DTW is very computationally expensive.

NCC is a shift-invariant similarity measure that is quicker to compute by defining the similarity through cross-correlation. Cross-correlation computes the similarity of two time series by slides one over the other and considering all possible positions for computing the inner product between them. It is then normalized to form the NCC. Then, the maximum value of all positions is taken to be the similarity between the two sequences. Normally, this computation takes $O(T^2)$, where T is the length of the sequences. However, since these computations are convolutions, Fast Fourier Transforms can be utilized to optimize the calculations to $O(Tlog(T))$. Details of the similarity measure can be referred to in [13].

To produce the input features through soft clustering, we consider the weekday daily load profile and weekend daily load profile of customers. We take Monday through Thursday as weekday, and Friday through Sunday as weekend. The daily profiles is obtained by averaging the daily load profiles of each customer on weekdays and weekends respectively by DTW barycenter averaging [15]. The profiles are then normalized through standardization. Finally, the output features are the membership vectors from clustering the weekday and weekend daily profiles, where the membership vector for a customer is the vector of similarity of the daily load profile of that customer to each cluster centroid.

Input Feature Extraction Similar to [7], instead of using consumption time series directly as an input, we extract a number of features related to consumption figures of the time series and ratio of these consumption figures. We define time periods morning, evening, night, noon as 6-10am, 6-10pm, 1-5am, 10am-2pm respectively. For consumption figure features, we take the mean, maximum (max) and minimum (min) of each time period over the full dataset (full), over weekdays (wd) only or over weekends (we) only. Table 1 summarizes the input features extracted with the feature names denoted by the abbreviations defined.

Table 1. List of consumption figure input features

Feature name	**Description**
full_mean	Mean consumption of the full dataset
full_max	Maximum consumption of the full dataset
full_min	Minimum consumption of the full dataset
wd_mean	Mean consumption on weekdays
wd_max	Maximum consumption on weekdays
wd_min	Minimum consumption on weekdays
we_mean	Mean consumption on weekends
we_max	Maximum consumption on weekends
we_min	Minimum consumption on weekends
full_morning_mean	Mean consumption during morning hours of the full dataset
wd_morning_mean	Mean consumption during morning hours on weekdays
we_morning_mean	Mean consumption during morning hours on weekends
full_evening_mean	Mean consumption during evening hours of the full dataset
wd_evening_mean	Mean consumption during evening hours on weekdays
we_evening_mean	Mean consumption during evening hours on weekends
full_night_mean	Mean consumption during night hours of the full dataset
wd_night_mean	Mean consumption during night hours on weekdays
we_night_mean	Mean consumption during night hours on weekends
full_noon_mean	Mean consumption during noon hours of the full dataset
wd_noon_mean	Mean consumption during noon hours on weekdays
we_noon_mean	Mean consumption during noon hours on weekends

Ratio features that are extracted are listed in Table 2 below.

Predictor After the feature extraction step, we need a predictor model that can map input features to target output feature values. Since we only consider binary value output features in this work, we can use a classifier as the model. In this paper, we pick the use of two different predictors: Support Vector Classifier (SVC) [16] and Feed-Forward Neural Network (NN).

SVC finds a hyper plane that separates input data points that should be mapped to different classes while maximizing the margin. SVC can only find a linear hyper plane which limits the modelling capability of the classifier. To

Table 2. List of ratio input features

full_mean/full_max	full_min/full_max
full_min/full_mean	wd_mean/wd_max
wd_min/wd_max	wd_min/wd_mean
we_mean/we_max	we_min/we_max
we_min/we_mean	full_morning_mean/full_noon_mean
full_evening_mean/full_noon_mean	wd_morning_mean/wd_noon_mean
wd_evening_mean/wd_noon_mean	we_morning_mean/we_noon_mean
we_evening_mean/we_noon_mean	wd_max/we_max
wd_min/we_min	wd_mean/we_mean
wd_morning_mean/we_morning_mean	wd_evening_mean/we_evening_mean
wd_night_mean/we_night_mean	wd_noon_mean/we_noon_mean

overcome this, we use a Radial Basis Function kernel [17] to introduce non-linearity into the classification process.

Feed-Forward NN is a neural network architecture that consists of only fully connected layers. In this paper, we use Rectified Linear Units (ReLU) as the activator function of the hidden layers and a sigmoid function as that of the output layer. The loss is computed by binary cross entropy loss as the output labels only takes on binary values.

5 Experiment and Results

Dataset and Features Extraction For our experiments, we use the dataset provided by the Irish Commission for Energy Regulation (CER) [4]. This dataset is a survey of customer electricity consumption behavior done in 2009. Consumption data is available for 4323 households from July 2009 to December 2010 in 30 minute intervals. In addition, a pre-trial and post-trial survey is carried out on each customer, which contains responses to a wide range of questions involving characteristics of the customers, their houses, and their living habits. Similar to [7], we take electricity consumption data for each customer from Week 2-5 (July 5th 2009 - August 1st 2009) for the experiments as they are closest to the trial survey.

We extract 17 different features from the Irish CER dataset related to households and their inhabitants. These are the target features to be predicted. The features picked are chosen with reference to the features used in other similar works [7, 6]. We focus mainly on three kinds of features:

1. Characteristics of the house: This includes features like the age of the house, the area of the house, etc.
2. Appliances in the house: We are concerned with certain kind of electric appliances that have high energy consumption, e.g. Electric Heater, Tumble Dryer, etc.
3. Characteristics of the inhabitants: This includes information regarding the main income earner and other residents in each household.

Table 3 lists all the features we have defined. We define each feature as a binary problem to simplify the prediction process. The last column in the table shows the definition of each binary problem. In some features, some customer may have refused to provide an answer, we remove these responses from the training and testing dataset for the considered feature only.

Table 3. List of features extracted

No.	Feature	Description	Binary Problem
1	Social Class	Social Class of the main income earner in the household	A,B,C **or** D,E
2	House Age	Age of the house (in years)	$\leq$30 **or** >30
3	Electric Cooker	Electric Cooker present in household	Yes **or** No
4	Electric Heater	Electric Heater present in household	Yes **or** No
5	House Floor Area	Floor Area of the house (in square feet)	<180 **or** $\geq$180
6	Children	Have children living in household	Yes **or** No
7	Resident Number	Number of residents living in household	$\leq$ 2 **or** >3
8	Residents in Daytime	Have residents living in household in daytime	Yes **or** No
9	Retirement	Main income earner in household has retired	Yes **or** No
10	Age	Age of main income earner in household	<65 **or** $\geq$65
11	Detached House	Is the house a detached house or connected house	Yes **or** No
12	Number of Bedrooms	Number of Bedrooms in the house	$\leq$2 **or** >3
13	Tumble Dryer	Tumble Dryer present in household	Yes **or** No
14	Dishwasher	Dishwasher present in household	Yes **or** No
15	Stand alone freezer	Stand alone freezer present in household	Yes **or** No
16	Education	Education level of main income earner in the household	less than primary **or** above second level
17	Income	Annual income of the main income earner in the household	$\leq$ 55k **or** >55k

Baselines We use two naive methods as baselines to compare against results from using extracted input features. These two methods are statistical methods that make prediction based on the statistics of the target feature values only:

- Identical Outputs (Identical) - Outputs either all “True” or all “False” for a feature. Choose “True” or “False” depending on whether the training dataset has more ”True” or ”False” values for that feature.
- Biased Random Guess (Biased) - Make a biased random guess based on how likely a value is “True” or “False” for a feature. Estimate the probability of a feature d having a “True” outcome $P(x_d = \text{“}T\text{”}) = \frac{N_d^{(T)}}{N}$ where $N_d^{(T)}$

is the number of “True” values in the training dataset for feature d and N is the total number of data points in the training dataset. Then, for each prediction for feature d in the testing dataset, output ”True” with probability $P(x_d = \text{“}T\text{”})$ or “False” with probability $1 - P(x_d = \text{“}T\text{”})$.

Table 4. List of input features configurations used for each experiment

No.	Input Features Used
1	Clusters only
2	Consumption and Ratio
3	Clusters, Consumption and Ratio

Experiment Setup and Evaluation Metrics In the experiments, we predict the target features using a Support Vector Classifier (SVC) model or a Feed Forward Neural Network (NN) with the prepared input features. In Section 5, we have described three types of input features: Consumption features, ratio features and cluster features. We test the performance of the SVC predictor when using different sets of features. Table 4 summarizes the 3 set of input features combination we use. We label each method the model initials and the input features configuration number, for example, ”SVC1” denotes using SVC model with cluster features only, ”NN3” denotes using NN model with all input features.

Since all target features are binary values, the result prediction can either be:

- True Positive (TP): Predicting a ”True” result correctly
- True Negative (TN): Predicting a ”False” result correctly
- False Positive (FP): Predicting a ”True” result as a ”False”
- False Negative (FN): Predicting a ”False” result as a ”True”

We evaluate the results by computing the classification accuracy and balanced accuracy as used by [6]. The classification accuracy simply calculates the rate of correct predictions, while the balanced accuracy is the average of correct ”True” predictions and correct ”False” predictions. The balanced accuracy is able to better show the ability of the predictor to predict correctly when the target feature is heavily biased towards either ”True” or ”False”.

$$\text{Classification accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Balanced accuracy} = \frac{1}{2} \times \frac{TP}{TP + FN} + \frac{1}{2} \times \frac{TN}{TN + FP} \tag{2}$$

The dataset of 4323 customers is split into training, validation and testing dataset by a $4 : 1 : 5$ split. The model is trained by the training dataset and validation is used to search for the best hyperparameters for the predictor model. The results are evaluated by the classification accuracy and balanced accuracy on the testing dataset.

Table 5. Classification accuracy for each input feature configuration with SVC predictor compared against baselines

No	Feature	Identical	Biased	SVC1	SVC2	SVC3	MissRate(%)
1	Social Class	60.92	52.94	60.87	63.37	63.12	96.50
2	House Age	56.24	49.57	54.63	57.42	58.18	100.00
3	Electric Cooker	74.79	64.57	74.79	75.50	76.30	99.91
4	Electric Heater	69.87	58.70	69.87	69.82	69.82	99.91
5	House Floor Area	96.55	93.77	96.55	96.55	96.55	42.49
6	Children	71.31	59.26	71.31	71.88	71.60	100.00
7	Resident Number	71.83	58.36	71.83	72.02	72.07	100.00
8	Residents in Daytime	65.08	57.18	68.38	71.79	71.93	100.00
9	Retirement	68.86	57.37	68.86	69.38	69.80	100.00
10	Age	77.08	66.40	77.08	77.13	77.08	100.00
11	Detached House	83.40	73.94	83.40	83.40	83.40	99.91
12	Number of Bedrooms	90.73	82.07	90.73	90.68	90.68	99.91
13	Tumble Dryer	66.97	59.26	69.47	71.08	70.60	100.00
14	Dishwasher	66.21	57.23	68.95	71.08	70.94	100.00
15	Stand alone freezer	50.43	49.95	56.85	63.61	64.08	100.00
16	Education	86.65	76.15	86.65	86.65	86.65	94.52
17	Income	50.67	49.64	50.98	52.02	54.20	45.60
-	Average (Mean)	71.03	62.73	71.84	73.14	73.35	92.87

Results and Discussion Table 5 and 6 shows the classification accuracy and balanced accuracy of the baselines and SVC using the different input configurations. The final column shows the MissRate, which is the rate of missing data for each feature in the testing dataset. We could observe that not all features are correlated to features extracted from consumption data. For "Electric Heater", "House Floor Area", "Age", "Detached House", "Number of Bedrooms", and "Education", the results were either the same as "Identical" outputs baseline or were within a very small margin of less than 0.1% difference. We could also see that the balanced accuracy for these features were 50.00% for all baselines as they classified all features correctly for one class and all incorrectly for the other. This meant that the SVC was not able to find a good classification for the input features and simply output the same class for all samples. This may be due to these features being not correlated to the electricity consumption. For "Electric Heater" in particular, since only summer electricity consumption data was used in this experiment, the heater may not have been used in this period.

Next, we compare the performance of "SVC3", which includes all input features including clustering, consumption figures and ratio features, to "SVC2"

Table 6. Balanced accuracy for each input feature configuration with SVC predictor compared against baselines

No	Feature	Identical	Biased	SVC1	SVC2	SVC3	MissRate(%)
1	Social Class	50.00	50.68	50.05	57.04	56.89	96.50
2	House Age	50.00	49.20	53.60	56.69	57.41	100.00
3	Electric Cooker	50.00	51.19	50.00	52.28	54.06	99.91
4	Electric Heater	50.00	50.72	50.00	50.01	50.01	99.91
5	House Floor Area	50.00	48.56	50.00	50.00	50.00	42.49
6	Children	50.00	50.32	50.00	53.45	53.25	100.00
7	Resident Number	50.00	48.22	50.00	54.26	54.35	100.00
8	Residents in Daytime	50.00	51.77	56.93	61.99	62.51	100.00
9	Retirement	50.00	49.35	50.00	52.12	53.47	100.00
10	Age	50.00	51.69	50.00	50.18	50.22	100.00
11	Detached House	50.00	51.29	50.00	50.00	50.00	99.91
12	Number of Bedrooms	50.00	48.89	50.00	49.97	49.97	99.91
13	Tumble Dryer	50.00	51.90	56.33	60.07	59.42	100.00
14	Dishwasher	50.00	51.27	56.42	61.00	60.69	100.00
15	Stand alone freezer	50.00	49.96	56.92	63.62	64.10	100.00
16	Education	50.00	49.80	50.00	50.00	50.00	94.52
17	Income	50.00	49.60	50.56	51.79	53.98	45.60
-	Average (Mean)	50.00	50.26	51.81	54.38	54.72	92.87

which does not contain clustering features. In particular, ”SVC3” performed better than ”SVC2” for ”House Age”, ”Electric Cooker”, ”Residents in Daytime”, ”Retirement”, ”Stand Alone Freezer” and ”Income”. This shows that the addition of clustering features has aided characteristic predictions for characteristics involving mainly resident occupancy, electric appliances with frequent usage like electric cooker and freezer, and characteristics concerning the income of the household like the annual income and the retirement status. Among only these features, the average accuracy of ”SVC3” was 65.75% while that of ”SVC2” is only 64.95%. On the other hand, ”SVC3” performed worse than ”SVC2” for ”Social Class”, ”Children”, ”Tumble Dryer” and ”Dishwasher” feature and did not see improvements for features like ”Age” and ”Education”, showing that clustering features did not provide much information about personal characteristics involving the main income earner. Among only these features, the average accuracy of ”SVC3” was 69.07% while that of ”SVC2” was 69.35%. The difference is smaller compared to the average accuracy on features where ”SVC3” performs better. On average, we could see that ”SVC3” has the highest classification accuracy of 73.35% among all features compared to ”SVC2” with 73.14%.

Table 7 and 8 shows the classification accuracy and balanced accuracy of using NN as the prediction model instead of SVC. Similar to SVC, ”NN3” which used all input features performed better than ”NN1” and ”NN2” which only used either cluster features or statistical features. The overall accuracy of ”NN3” however was lower than ”SVC3”. This was compensated by the fact that ”NN3” was consistently performing better than ”NN2” in most features. ”NN3” only had a

Table 7. Classification accuracy for each input feature configuration with NN predictor compared against baselines

No	Feature	Identical	Biased	NN1	NN2	NN3	MissRate(%)
1	Social Class	60.92	53.09	60.48	62.14	62.83	96.50
2	House Age	56.24	49.67	55.01	58.22	58.41	100.00
3	Electric Cooker	74.79	62.58	74.79	76.02	76.11	99.91
4	Electric Heater	69.87	58.04	69.87	69.87	69.73	99.91
5	House Floor Area	96.55	94.10	96.55	96.55	96.55	42.49
6	Children	71.31	56.00	71.31	71.50	72.02	100.00
7	Resident Number	71.83	59.17	71.83	72.40	72.35	100.00
8	Residents in Daytime	65.08	54.54	68.19	73.16	73.49	100.00
9	Retirement	68.86	57.56	68.86	69.23	68.95	100.00
10	Age	77.08	64.13	77.08	77.41	77.17	100.00
11	Detached House	83.40	72.71	83.40	83.40	83.40	99.91
12	Number of Bedrooms	90.73	81.79	90.73	90.73	90.73	99.91
13	Tumble Dryer	66.97	56.00	69.38	71.41	71.46	100.00
14	Dishwasher	66.21	53.78	68.81	71.08	71.12	100.00
15	Stand alone freezer	50.43	48.91	58.03	63.23	63.23	100.00
16	Education	86.65	77.00	86.65	86.65	86.65	94.52
17	Income	50.67	48.81	50.67	50.67	50.67	45.60
-	Average (Mean)	71.03	61.64	71.86	73.16	73.23	92.87

lower classification accuracy than ”NN2” in ”Electric Heater”, ”Resident Number”, ”Retirement”, and ”Age”. On the other hand, ”NN3” had higher classification accuracy than ”NN2” in ”Social Class”, ”House Age”, ”Electric Cooker”, ”Children”, ”Residents in Daytime”, ”Tumble Dryer”, and ”Dishwasher”. For the remaining features, no improvements can be seen using an NN prediction model compared to the ”Identical” baseline.

6 Future Work

In this paper, we used a similarity based soft clustering approach. This means that the the clusters are generated based on a particular similarity measure defined for the consumption time series. While this approach can perform well if the similarity measure is properly defined, it has limited modelling power to capture electricity consumption habits of customers which can be complicated. There are many hidden latent factors that affect the electricity usage patterns of a customer, including the day of week, temperature, habitual behavior, etc. In our methodology, we took one of these factors, day of week, into consideration by separately clustering the daily load profiles of weekdays and weekends. However, manually modelling all such factors in this manner is not possible because many of these factors are not known, and it would lead to formation of many different clustering results. Instead, we can make use of generative models for the profiling of each customer. For example, the authors in [18] use a Dirichlet process mixture model to model the distribution of consumption load profiles. The membership

Table 8. Balanced accuracy for each input feature configuration with NN predictor compared against baselines

No	Feature	Identical	Biased	NN1	NN2	NN3	MissRate(%)
1	Social Class	50.00	50.85	52.63	60.37	61.43	96.50
2	House Age	50.00	49.24	53.29	57.86	58.11	100.00
3	Electric Cooker	50.00	48.37	50.00	54.24	54.80	99.91
4	Electric Heater	50.00	49.84	50.00	50.62	50.08	99.91
5	House Floor Area	50.00	53.40	50.00	50.00	50.00	42.49
6	Children	50.00	47.59	50.00	56.14	56.16	100.00
7	Resident Number	50.00	49.50	50.00	57.89	57.96	100.00
8	Residents in Daytime	50.00	49.21	56.13	64.52	65.27	100.00
9	Retirement	50.00	49.74	50.00	53.06	53.06	100.00
10	Age	50.00	49.57	50.00	51.95	51.87	100.00
11	Detached House	50.00	50.09	50.00	50.00	50.00	99.91
12	Number of Bedrooms	50.00	49.20	50.00	50.23	50.00	99.91
13	Tumble Dryer	50.00	49.21	55.53	61.65	61.76	100.00
14	Dishwasher	50.00	48.25	56.17	62.72	62.89	100.00
15	Stand alone freezer	50.00	48.91	58.05	63.21	63.22	100.00
16	Education	50.00	50.45	50.00	50.00	50.00	94.52
17	Income	50.00	48.74	50.00	50.00	50.00	45.60
-	Average (Mean)	50.00	49.54	51.87	55.56	55.68	92.87

vector of each customer can be represented as the probability of their observed daily load profile being generated by each Dirichlet process model.

7 Conclusion

In this paper, we proposed the use of soft clustering results to complement electricity consumption related statistical features as inputs for socio-demographic characteristics prediction. We performed soft clustering on weekday and weekend daily load profiles of customers using SImilarity-Based Soft Clustering (SISC) [12] with Normalized Cross Correlation (NCC) [13] as the similarty measure. The membership vector of each customer which represents the similarity between their load profile to each cluster was used as input features for characteristics prediction.

We experimented the performance of socio-demographic characteristics prediction using a real life dataset with customer characteristics partially collected through survey. We handpicked 17 household characteristics that covered information related to the house, the electricity appliances and the inhabitants. The characteristics were defined as binary value problems and a prediction model was trained to predict each characteristics from features extracted from load profiles. We compared the prediction accuracy using two different prediction models: Support Vector Classifier and Feed-Forward Neural Network. Experiments showed that overall, the classification accuracy for characteristics prediction was higher when both soft clustering features and statistical features were used as inputs than solely using either set of features. This showed that the inclusion of soft clustering features improved predictions for certain characteristics that were highly correlated to the clustering results.

8 Acknowledgements

This work has been sponsored by the U.S. Army Research Office (ARO) under award number W911NF1910362 and the U.S. National Science Foundation (NSF) under award numbers 1911229 and 2009057.

Bibliography

[1] Joaquim L Viegas, Susana M Vieira, João MC Sousa, R Melicio, and VMF Mendes. Electricity demand profile prediction based on household characteristics. In *2015 12th International Conference on the European Energy Market (EEM)*, pages 1–5. IEEE, 2015.

[2] Peter Palensky and Dietmar Dietrich. Demand side management: Demand response, intelligent energy systems, and smart loads. *IEEE Transactions on Industrial Informatics*, 7(3):381–388, 2011.

[3] Ailin Asadinejad, Alireza Rahimpour, Kevin Tomsovic, Hairong Qi, and Chien-fei Chen. Evaluation of residential customer elasticity for incentive based demand response programs. *Electric Power Systems Research*, 158:26–36, 2018.

[4] Cer smart metering project - electricity customer behaviour trial, 2009-2010. accessed via the irish social science data archive - www.ucd.ie/issda, 2009-2010.

[5] Danielly B Avancini, Joel JPC Rodrigues, Simion GB Martins, Ricardo AL Rabêlo, Jalal Al-Muhtadi, and Petar Solic. Energy meters evolution in smart grids: A review. *Journal of cleaner production*, 217:702–715, 2019.

[6] Joaquim L Viegas, Susana M Vieira, and João MC Sousa. Fuzzy clustering and prediction of electricity demand based on household characteristics. In *2015 Conference of the International Fuzzy Systems Association and the European Society for Fuzzy Logic and Technology (IFSA-EUSFLAT-15)*. Atlantis Press, 2015.

[7] Gan Sun, Yang Cong, Dongdong Hou, Huijie Fan, Xiaowei Xu, and Haibin Yu. Joint household characteristic prediction via smart meter data. *IEEE Transactions on Smart Grid*, 10(2):1834–1844, 2017.

[8] Michel Verleysen and Damien François. The curse of dimensionality in data mining and time series prediction. In *International work-conference on artificial neural networks*, pages 758–770. Springer, 2005.

[9] Christian Beckel, Leyna Sadamori, and Silvia Santini. Automatic socio-economic classification of households using electricity consumption data. In *Proceedings of the fourth international conference on Future energy systems*, pages 75–86, 2013.

[10] Marta P Fernandes, Joaquim L Viegas, Susana M Vieira, and João M Sousa. Analysis of residential natural gas consumers using fuzzy c-means clustering. In *2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1484–1491. IEEE, 2016.

[11] Seunghyoung Ryu, Minsoo Kim, and Hongseok Kim. Denoising autoencoder-based missing value imputation for smart meters. *IEEE Access*, 8:40656–40666, 2020.

[12] King-Ip Lin and Ravikumar Kondadadi. A similarity-based soft clustering algorithm for documents. In *Proceedings Seventh International Conference*

on Database Systems for Advanced Applications. DASFAA 2001, pages 40–47. IEEE, 2001.

[13] John Paparrizos and Luis Gravano. k-shape: Efficient and accurate clustering of time series. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 1855–1870, 2015.

[14] Meinard Müller. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84, 2007.

[15] François Petitjean, Alain Ketterlin, and Pierre Gançarski. A global averaging method for dynamic time warping, with applications to clustering. *Pattern recognition*, 44(3):678–693, 2011.

[16] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[17] Bernhard Scholkopf, Kah-Kay Sung, Christopher JC Burges, Federico Girosi, Partha Niyogi, Tomaso Poggio, and Vladimir Vapnik. Comparing support vector machines with gaussian kernels to radial basis function classifiers. *IEEE transactions on Signal Processing*, 45(11):2758–2765, 1997.

[18] Ramon Granell, Colin J Axon, and David CH Wallom. Clustering disaggregated load profiles using a dirichlet process mixture model. *Energy Conversion and Management*, 92:507–516, 2015.